

THEORY AND APPLICATION OF NEAREST NEIGHBOR IMPUTATION IN CENSUS 2000

Robert E. Fay*

U.S. Census Bureau, Washington, DC 20233-9001

Key words: missing data, variance estimation, replication, replicate weights.

1. Introduction

1.1 Changing Census Plans The evolving plans for the next decennial census in the U.S., Census 2000, have been widely reported in the press and remain a focus of political debate. Last year, Thompson and Fay (1998) reviewed the milestones in the development of the decennial program at the Joint Statistical Meetings, particularly with respect to statistical sampling and estimation, which was to have been one of the cornerstones of the plans. While the 1970 census provided a precedent for inclusion of sampling and estimation in producing the population count (Wright 1999), Census 2000 was to have been the first in the U.S. designed with the goal of achieving an optimal combination of counting, assignment, and estimation in order to obtain population totals.

The 1980 and 1990 censuses primarily employed a mail strategy in which households in most parts of the country were mailed or delivered census forms. The majority of households responded by mail, but nonresponding households were followed up by personal visit to complete the enumeration. In some cases, interviewers had to obtain information about nonrespondents from neighbors or other sources, although precise figures on such proxy response are unavailable.

Both demographic analysis (Robinson, Ahmed, Das Gupta, and Woodrow 1993) and coverage studies using sample surveys have documented a persistent undercount of some groups, including Blacks. In both 1980 and 1990, the issue of a potential adjustment to the census counts to compensate for differential undercoverage became a matter of both debate and litigation. Results from coverage evaluations could not be produced until months after the release of the apportionment counts used to allocate the number of representatives among the states. No official numbers were adjusted in 1980. In 1990, Secretary of Commerce Mosbacher decided in July, 1991, against the Census Bureau recommendation for adjustment. The only official figures incorporating an adjustment for undercount in the 1990 census are postcensal population estimates at the state and national level used as controls for some demographic surveys, such as the Current Population Survey (CPS), the monthly labor force survey in the U.S. In particular, the official postcensal population estimates, published and

used for other purposes including allocation of funds, do not incorporate an adjustment for census undercount, unlike current Canadian practice (Germain and Julian 1993, Dick 1995).

As of August 1998, the Census Bureau's plan for Census 2000, developed over several years, employed statistical sampling and estimation in two primary ways (Thompson and Fay 1998):

- 1) A sample of nonresponding households was to be selected for Nonresponse Followup (NRFU), and the results used to form estimates for nonsample nonresponding households. Sampling for NRFU offered savings in both time and money. Nonetheless, this sample was to have been extremely large, on the order of tens of millions of households, enabling conventional survey estimates down to very low levels of geography.
- 2) A large coverage study, called the Integrated Coverage Measurement (ICM), would be based on approximately 750,000 housing units and viewed as an integral part of the census. Its purpose was to measure differential undercoverage. The results would be incorporated into all official results, including the state population counts delivered to the President on December 31, 2000, to be used for the apportionment of the U.S. House of Representatives among the 50 states.

In January, 1999, however, the U.S. Supreme Court upheld lower court rulings that the current census legislation did not permit the use of either form of sampling for the apportionment. The court's ruling did not resolve the constitutionality of such a census if current law were revised to permit it. Plans for Census 2000, however, were changed to conform with the existing legislation and the court's interpretation of it.

Thus, the effect of the Supreme Court ruling was to eliminate the first type of sampling, sampling for NRFU, from Census 2000. In other words, once NRFU sampling is excluded for purpose of apportionment, it would have no practical use for any other purpose. Revised plans for the census now reflect the increased workload and time requirements to follow up roughly 15,000,000 more nonresponding housing units. State counts without the use of sampling and estimation will be produced by December 31, 2000, for the apportionment.

Although NRFU sampling was entirely eliminated by the court's decision, the ICM has been redesigned as the Accuracy and Coverage Evaluation (A.C.E.). Indeed, the

court's ruling did not exclude the use of sampling for uses other than apportionment, and to some degree suggested that sampling would be appropriate if feasible. The A.C.E. sample size has been reduced from 750,000 to approximately 300,000 housing units, and the time schedule has been adjusted for the more extensive period required for NRFU. Current plans are to obtain A.C.E. estimates by approximately February 15, 2001, relatively earlier than previous studies in 1980 and 1990. The A.C.E. estimates, in the form of estimated percent undercounts for a set of poststrata, are planned to be incorporated into the detailed official counts down to the block-level for release by April 1, 2001. This timing permits their potential use in redistricting. (In the U.S., states allocated more than one representative in the House of Representatives are divided into Congressional Districts on the basis of population. *Redistricting* is the process of defining these boundaries. Although court rulings now tightly limit variation in the population sizes of congressional districts within a state, states are given some latitude in other respects in determining their boundaries.) Under the revised plan, all official data products from Census 2000, except the apportionment counts, will incorporate the results of the A.C.E.

1.2 Nearest Neighbor Imputation in the Dress Rehearsal The paper will focus on the nearest neighbor imputation as an estimation procedure for NRFU in the Census 2000 Dress Rehearsal in Sacramento and on an associated variance estimator. Thus, the paper concerns methodological aspects of an application obviated by the Supreme Court's ruling. Nonetheless, this paper, and one in preparation (Fay and Farber 1999), will focus on methodological findings from the Dress Rehearsal effort.

The 2000 Dress Rehearsal was conducted in three sites. Through agreement with Congress, the Census Bureau implemented its plan for Census 2000, combining sampling for NRFU with Integrated Coverage Measurement, only in Sacramento, California. A site in Columbia, South Carolina, and surrounding counties was enumerated without the statistical methods, although an accompanying Post-Enumeration Survey, similar in design to the ICM, was employed as an evaluation. A smaller site in Wisconsin did not use sampling for NRFU but incorporated the ICM corrections.

The selection of a probability sample for NRFU theoretically permits the use of standard survey weighting procedures. Instead, a nearest neighbor/hot-deck imputation method was implemented in the Sacramento site, in effect treating nonsample nonresponding housing units as a problem in unit nonresponse. The subject of section 2 and a principal focus of this paper is the rationale for this methodological selection. The section also describes the Sacramento application in more detail.

1.3 Variance Estimation A second purpose of the paper is to present a suitable variance estimator for the NRFU imputation. The variance estimator has potential application to other nearest neighbor imputation situations.

As often noted, the term *hot-deck imputation* has been applied to a variety of similar methodologies. For purposes of discussion in this paper, it is useful to group most of these into three broad categories:

1. *The sequential hot deck.* This original form appears to have been substantially shaped by available computer resources and practice at the time of its development. In its simplest form, a characteristic x is available for all units but y is subject to possible nonresponse. Units are classified on the basis of x into prespecified cells. Typically, an array is loaded with a *cold deck* of initial values based on an earlier survey or some other suitable source. The units are processed sequentially, often in a sort reflecting geographic proximity or another measure of similarity. New units with observed y , termed *donors*, are used to replace old values in the hot-deck array; units with missing y are assigned values from the hot deck. For example, the empirical study of Rizvi (1983) considered only this form of hot deck, whereas the overview by Ford (1983) considered this form of hot deck as well as nearest neighbor imputation below.
2. *Statistical matching with fixed cells.* The ability to sort moderate or large files and other forms of data access removed the restriction that the hot deck be tied to the sequential order of the data file. For example, the variance estimator developed by Rao and Shao (1992) was for a hot deck unconstrained by the order of the file. Specifically, they consider units cross-classified into a potentially large number of cells, and each unit requiring imputation can receive a value from any of the donors falling in the corresponding cell. In Rao and Shao (1992), observations requiring an imputation are independently assigned values with probability proportional to each donor's survey weight. The asymptotic argument in Rao and Shao (1992) permitted an increasing number of cells but required that generally each cell have an increasing number of eligible donor cases. Fay (1996) contrasted the Rao and Shao (1992) estimator for this situation with a multiple imputation (Rubin 1987) approach, finding a clearer frequentist interpretation for the Rao and Shao approach than for multiple imputation.
3. *Nearest neighbor imputation.* This form extends the logic of statistical matching further, searching for either a unique best match or small number of

equivalent matches on the basis of x among units with observed y . The cells in this form of imputation are not predetermined. For example, Colledge, Johnson, Paré, and Sande (1978) describe an application to economic data in which a specific set of nearest neighbors were identified for each case requiring imputation. I.G. Sande (1983) reviewed the general features of the nearest neighbor approach, and G.T. Sande (1983) provided additional comments on its features and computational feasibility. Lee, Rancourt, and Särndal (1994) presented and studied a variance estimator for y using nearest neighbor imputation on the basis of a continuous x and an assumed regression model for y on x . The extension of multiple imputation to this form of nearest neighbor imputation is less clear, and the version of multiple imputation investigated by Lee, Rancourt, and Särndal did not perform as satisfactorily in an empirical study as their own estimator.

This grouping is useful for categorizing variance estimators for the hot deck, but the distinctions among the three groups are not always precise, particularly between the second and third form. Bankier, Luc, Nadeau, and Newcombe (1996) characterize the New Imputation Methodology (NIM) developed at Statistics Canada (Bankier, Houle, Luc, and Newcombe 1997) as a minimum change hot deck, and it appears appropriate to categorize the NIM in the third group. The current imputation methodology for work history and income items for the Annual Demographic Supplement ("March Supplement") to the Current Population Survey (CPS) in the U.S. employs a series of sorts to achieve a statistical match between donors and cases requiring imputation. (Coder 1978 and David, Little, Samuhel, and Triest 1986 provide a more detailed summary of the algorithm.) In effect, several different cross-classifications of observed characteristics are considered for each case requiring imputation. In many cases, a large number of donors may be available for particular cases requiring imputation, but in others cases imputations may be made by selection from a relatively small set of donors. Thus, the CPS application is best categorized as belonging to the third group but having ties to the second. Other versions of the hot deck, such as the use of the nearest neighbor approach with distance determined in whole or in part by predicted values from a parametric model, do not fit neatly into any of the three groups.

Fay developed a variance estimator especially for the third group, based on different assumptions than used by Rao and Shao. Unlike the variance estimator proposed by Lee, Rancourt and Särndal, the method does not require a parametric model. A key feature of the

method is the use of data from a second nearest neighbor for purposes of variance estimation. Various forms of the variance estimator have appeared earlier (Town and Fay 1995, Steel and Fay 1995, Fay and Town 1996, 1998). Section 3 describes the rationale of the estimator in more detail.

2 Selection of Nearest Neighbor Imputation in the Dress Rehearsal

2.1 Methodological Background

2.1.1 Block vs. Unit Sampling The Census Bureau's planned use of sampling for both NRFU and the correction of differential census undercoverage through ICM led to a number of separate research efforts on how to estimate census results under NRFU sampling. The design for NRFU sampling was constrained, however, to be consistent with the plans for the ICM.

The design strategy for the ICM involved sampling census blocks or block clusters and typically including all of the housing units in the sampled blocks in the ICM sample. (The largest blocks were to be subsampled. The average size of ICM clusters was projected to be about 30 housing units after subsampling.) Because of the required matching of ICM sample cases to their initial census counterparts, the design called for 100% NRFU, rather than NRFU sampling in ICM blocks. NRFU sampling in ICM blocks would have induced complexities arising from attempting to match an ICM household to a nonsample household in NRFU.

Two primary candidate designs were available for NRFU sampling of non-ICM blocks:

- Block sampling, where a sample of blocks with mail nonresponse would be selected. All nonresponding housing units in sampled NRFU blocks would be followed up.
- Unit sampling of nonresponding housing units in an unclustered manner. One form, implemented in one of two panels in Oakland, CA, in the 1995 Census Test, assured selection of sample units in all blocks with nonresponse. (For example, a single nonresponding unit in a block was always included.) In Sacramento, a systematic sample of nonresponding units was selected, resulting in sampled NRFU units in most but not all blocks with nonresponse.

The block-based design logically fit better with the ICM design, since NRFU would then be completed on a block basis in both ICM and non-ICM blocks. In other words, if

$$E(y \mid \text{unit sampling}) = E(y \mid \text{block sampling})$$

then statistical adjustment of counts from NRFU using unit sampling on the basis of an ICM using block sampling would be problematic. A distinct potential disadvantage, however, later confirmed by empirical studies, was that sampling blocks for NRFU could yield much higher variances.

Initially, research efforts had concentrated on NRFU estimation under block sampling. Tsay, Isaki, and Fuller (1996) and Zanutto and Zaslavsky (1996) investigated block-level models. Estimates at the block level would then be used as constraints by statistical procedures to estimate the nonsampled households and persons within each block to be added to the results of direct enumeration. Schaefer (1995) investigated a different procedure, which modeled at the household and person level. In Schaefer's approach, estimates for blocks emerged by summation. Schaefer's approach also attempted to integrate estimation for NRFU with imputation for item nonresponse for data on characteristics.

In the 1995 Census Tests, a split panel experiment compared unit and block sampling designs for NRFU sampling in Oakland, CA. The two panels produced post-NRFU estimates within sampling error of each other, but variances were substantially less for unit sampling (Fay and Town 1996). In other words, the evidence indicated

$$E(y \mid \text{unit sampling}) \approx E(y \mid \text{block sampling})$$

$$\text{Var}(y \mid \text{unit sampling}) \ll \text{Var}(y \mid \text{block sampling})$$

In general, cost is an additional factor to consider in selecting among survey designs. In this respect, block sampling was thought to have a small advantage in terms of slightly reduced travel, but the differences were considered marginal because both designs, with their high sampling fractions, would be very densely distributed.

The findings became the impetus to move to unit sampling. A related study based on matching ICM blocks to similar non-ICM blocks for the 1998 Dress Rehearsal showed no systematic differences (U.S. Census Bureau 1999a).

The effect of unit sampling is suggested by the following example: for a typical block of 30 housing units, a 70% response rate leaves 9 nonresponding units, of which an expected 6 would be sampled for NRFU, leaving an expected 3 to be estimated. Use of sampling rates defined within each block, or, as in the Dress Rehearsal, systematic sampling, are available to limit random variation in the realized sample size within blocks. Although the previous research had shown that models can make effective use of block-level variables in

forming estimates under block sampling, unit sampling reduces these gains by largely eliminating the effect of between-block variability on the estimate.

2.1.2 Weighting vs. Imputation With the selection of unit sampling and the accompanying feasibility of using simpler design-based estimators, a traditional weighting approach appeared to deserve consideration. Although the design-based rationale for weighting is clear, there were a few significant constraints. For the sake of ease of use, weights could only be integers. (Indeed, the Census Bureau has consistently used integer weights even for the census "long-form" sample data. In Census 2000, sample data will be collected from approximately 1 out of 6 households.) For NRFU, most weights would thus be either 1 or 2, except in areas where the response rate exceeded 80%.

For higher levels of geography such as census tracts, application of integer weights would have provided estimates of population plausibly consistent with housing unit totals. Use of weights at the block level, however, could have produced marked inconsistencies between population counts directly affected by NRFU sampling and counts of housing units, which were almost free from random variation.

Use of nearest neighbor imputation, with nonsampled nonresponse housing units imputed from nearby sampled NRFU units, maintained the integer nature of the census data and linkage between housing units and population. Farber and Griffin (1998) compared weighting vs. nearest neighbor imputation, and found almost equal performance at most geographic levels, but awarded the advantage to nearest neighbor imputation for its ready handling of blocks lacking sampled NRFU units.

2.2 Basic Implementation Strategy in the Sacramento Dress Rehearsal

As in previous U.S. censuses, part of the overall task of taking a census is to obtain an inventory of housing units. Sampling was not to be employed at this stage; rather, the Master Address File (MAF) of all housing units was the frame for sampling. (Sampling enters slightly, in that the result of NRFU is occasionally to determine that a unit in the MAF must be deleted. For example, the unit might be a commercial address or demolished. Consequently, the imputation also classified some nonsample units to delete status.)

Sampling for NRFU was to be used for two types of incomplete data. The first of these was housing units not responding by mail: nonresponding housing units represent a mixture of households not returning their forms, vacant units, and units that should have been deleted from the MAF. Consequently, NRFU was to determine the occupancy status of sampled nonresponding units as well as other characteristics. The

sampling rates were set so that statistical estimation would be used to represent only 10% (or less) of the housing units in each census tract (a publication area generally representing about 3000 housing units). With mail nonresponse in Sacramento around 50%, overall about 4-in-5 nonresponding units were sampled for NRFU -- a very large sample compared to usual statistical practice.

The second type of incomplete data was entirely distinct from the first. During the mail delivery of the census questionnaires, carriers were allowed to return forms to the Census Bureau with the designation "Undeliverable as Addressed-Vacant" (UAA-vacant). From past experience, the majority of such units are vacant, but not all of them. (This again proved to be the case in Sacramento.) A sample of 3-in-10 of the UAA-vacant units was selected for followup visit by an enumerator, and units found to be occupied were enumerated. This sampling rate required that most UAA sample cases be used in imputation 2 or 3 times.

Each of the two groups was treated as an independent estimation problem, since the respective universes did not overlap. The following discussion will be primarily in terms of NRFU, although parallel operations were carried out for UAA. Farber, Fay, and Schindler (1998) describe the sampling and estimation in more detail, but a summary will be included here.

Based on the overall approach of section 2.1, the basic strategy was to identify a neighboring sample NRFU case as the basis for imputation for each nonsample nonresponse housing unit. In terms of the previous notation, characteristics x , including sort order in the MAF (which is related to location), census block, and basic address (generally enabling identification of units in the same building), are available for both sample and nonsample nonresponse units. Wherever possible, the matching algorithm selected donors from the same address for nonsample units at multi-unit addresses. Otherwise, the nearest neighbor of a nonsample case was defined on the basis of the distance when the file was sorted by block and MAF order within block. Thus, the algorithm favored using units in the same block, taking into consideration the closeness indicated by the MAF (U.S. Census Bureau 1999b). The overall design was that donor units would provide characteristics, y , including whether the unit should be deleted, occupancy status, and for occupied units, number and demographic characteristics of the residents, owner/renter status, etc.

In order not to depart from a design-based rationale too substantially, the number of times each eligible sample case was used was constrained to be consistent with the weight it would have had under a weighting approach. At the tract level, this weight would have been

$$w_{NRFU} = \frac{NRFU \text{ sample units} + NRFU \text{ nonsample units}}{NRFU \text{ sample units}}$$

omitting ICM blocks. A UAA weight, w_{UAA} , was defined similarly. For example, for a UAA sample case in a tract with $w_{UAA} = 3.33$, then the sample case was constrained to be used as a donor either 2 or 3 times (in addition to representing itself). Consequently, when the weights in a tract were an integer, then the nearest neighbor imputation was constrained to use each donor the same number of times and thus to reproduce at the census tract level the results of a weighting approach.

As a remark, had the Supreme Court ruled differently, and had sampling for NRFU remained part of Census 2000 plans, further empirical investigation of this approach would have been warranted. The procedure implemented for Dress Rehearsal insured a fairly high level of agreement between the imputation and weighting at aggregate levels, while effectively leaving the nearest neighbor approach some latitude in allocating whatever fractional weight was available. Because of high sampling rates for NRFU, in most tracts the weight did not reach 2, so the effect of the constraint was to allow the measure of distance to select donors but to constrain each donor to a single use at maximum. Alternative approaches, such as randomly determining which donors to use to distribute any fractional weight, would have provided even further protection against bias, possibly at the expense, however, of somewhat increased variance. Further empirical research on such points would have been warranted if NRFU sampling had remained in Census 2000 plans.

2.3 Modifications for Late Mail Returns

The conceptual model just described is based on a simple dichotomy between sample and nonsample nonresponse units. In practice, a cutoff date approximately 2 weeks after the due date was set during Dress Rehearsal, and the sample of nonresponding units was identified at that time. Some returns continued to be received thereafter. When a late mail return was received from a unit sampled for NRFU, a computer algorithm selected between the late return and the followup form that may have been obtained by personal visit on the basis of completeness, with a preference for the mail return. Late returns were incorporated into the census for nonsampled units, eliminating the need for any imputation for them. The decision to accept late returns was made on the basis of policy, to avoid rejection of

data from the public within the time span that the information could actually be processed and tabulated.

Superficially, late mail returns appear simply to reduce the size of the remaining followup somewhat. In fact, however, late mail returns cannot be regarded simply as a random sample from the outstanding units. Late mail returns, like mail returns in general, were almost exclusively from occupied housing units. Ignoring late returns, the nonsample nonrespondents are a probability sample of all nonrespondents as of the cutoff date, but when those returning late returns are removed, the remaining units should be somewhat skewed towards a population of vacants and deletes. Since the occupancy status belonged among the sample characteristics, y , rather than among the characteristics, x , available for matching, the use of nearest neighbor imputation would not correctly represent this aspect of response. In other words, although selecting a probability sample for NRFU would have led to nonresponse (where nonresponse in the general theory is equated to not a member of the NRFU sample in this instance) *unconfounded* with y , the subsequent exclusion of late mail returns from the nonsample cases leaves a set of nonsample cases with nonresponse *confounded* with y , particularly through occupancy status.

Lee, Rancourt, and Särndal (1994) investigated the performance of their proposed form of nearest neighbor imputation under both unconfounded and confounded response situations, but proposed no modification to the procedure in instances in which response is confounded.

A modification to the imputation to compensate for confounded response was implemented for the Dress Rehearsal. The approach was to reduce proportionately the number of available donors in accord with the number of late mail returns received for nonsample units. For each tract, a ratio of sampled to nonsampled units was computed; specifically (Farber, Fay, and Schindler 1998),

$$r_{NRFU} = \frac{\text{NRFU sampled addresses in non-ICM blocks}}{\text{NRFU nonsampled addresses in non-ICM blocks}} = \frac{1}{w_{NRFU} - 1}$$

Separately for the three categories, c , of sampled 1) occupied, 2) vacant, and 3) deleted units, a number of units,

$$Rm_c = r_{NRFU} \times LR_c = \frac{LR_c}{w_{NRFU} - 1}$$

were removed from the hot deck from c , where LR_c is the number of nonsampled late returns in c . Within c , sampled cases with late mail returns were targeted first for removal, followed by the remaining sample cases.

For $r_{NRFU} < 1$, that is, $w_{NRFU} > 2$, some sample units would be used more than once for imputation, and further research could have investigated an alternative that reduced the allowed imputations by 1 over a larger group of donors instead of completely eliminating some from any use. Again, circumstances no longer justify further pursuit of this option at this point.

2.4 NRFU/UAA Variances for Sacramento Although the official Dress Rehearsal result for Sacramento incorporated the ICM component, the Census Bureau has also released a population total without the ICM component of 377,741 for Sacramento. Using the methodology reported in the next section, the estimated standard error for this estimate was 321 people, or a c.v. of less than .09%. This standard error pertains to the effect of NRFU and UAA estimation. By contrast, when the ICM correction is also included, the official population total is 403,312 with estimated standard error 4,810 or a c.v. of 1.2%.

For a block of approximately 30 housing units and 75 people, a simple scaling of the site-level results suggests a standard error of approximately 4.5 people, or a c.v. of roughly 6%, from the NRFU/UAA component of estimation.

According to the ICM, the census count, 377,741, without the ICM for Sacramento is subject to an estimated 6.3% undercount (s.e. 1.1%).

2.5 Concluding Remarks on the Application Available evidence suggests that the implementation of nearest neighbor imputation essentially achieved its primary objective, namely, to produce statistical estimates comparable to conventional enumeration while agreeing closely in aggregate with traditional survey estimation results. Further empirical details will be provided by Fay and Farber (1999).

Changes in legislation could open the possibility of sampling for NRFU in 2010; if so, the 2000 Dress Rehearsal experience will represent a potential starting point for research efforts.

The application also illustrates consideration of the effect of confounded response in applying nearest

neighbor imputation. Although the procedure presented here could be further refined, the modification of the hot deck to compensate for the effects of confounded response may be a potential approach useful in other applications.

3 A Variance Estimator for Nearest Neighbor Imputation

3.1 Basic Rationale As noted in the first section, forms of the variance estimator had been reported in several previous collaborations. Although basic assumptions were stated, the previous accounts were considerably truncated. The discussion here is intended to provide an expanded account of the estimator. Section 3.1 will deal with a special case, which leads to a simplified, almost obvious, form. Section 3.2 provides the full estimator used in the Dress Rehearsal.

As noted in Section 1.3, the Rao-Shao variance estimator was developed relying on a random selection from a set of possible donors in the imputation cell. Forms of nearest neighbor imputation emphasizing use of an optimal nearest neighbor depart from conditions required for the Rao-Shao variance estimator.

For each case i requiring imputation, let y_i be the true but unobserved value and $y_i^* = y_{nn1(i)}$ the imputed value from the nearest neighbor, $nn1(i)$, of i . The scope of the variance estimator includes both applications in which the nearest neighbor, $nn1(i)$, is defined on the basis of distance only or is subject to constraints on the number of times each donor may be used, as in the Dress Rehearsal. The variance estimator employs the value from a second nearest (responding) neighbor, $nn2(i)$. In general, a second nearest neighbor, $nn2(i)$, for i may be defined by omitting the first nearest neighbor, $nn1(i)$, and applying the nearest neighbor definition/algorithm. (In the Dress Rehearsal application, the constraint on the number of uses of the donor was not employed in defining the second nearest neighbor, nor is it theoretically necessary to do so.)

The variance estimator is developed under a population model, ξ , for a population of y conditional on x (Fay and Town 1998). The variance estimator relies on model ξ assumptions:

$$E_{\xi}(y_i) = E_{\xi}(y_{nn1(i)}) = E_{\xi}(y_{nn2(i)}) \quad (1)$$

$$Var_{\xi}(y_i | x_i) = 1/2 E_{\xi}(y_{nn1(i)} - y_{nn2(i)})^2 \quad (2)$$

$$Cov_{\xi}(y_i, y_{i'} | x_i, x_{i'}) = 0, i \neq i' \quad (3)$$

To the extent that $x_{nn1(i)}$ and $x_{nn2(i)}$ are close to but not necessarily equal to x_i , assumption (1) represents an ideal that practical applications will generally only approximate. Assumption (3) follows from an assumption of independence of the y 's given the x . Assumption (2) follows from (1) and (3) if nonresponse is unconfounded with y given x .

Let Y denote the sum of the y 's in the given finite population, which is a realization under ξ . Suppose Y^* denotes the sum with the imputed values for y . If each nearest neighbor is used in imputation at most once, then,

$$\begin{aligned} E_{\xi}(Y - Y^*)^2 &= \sum_{i \in A_{\pi}} E_{\xi}(y_i - y_{nn1(i)})^2 \\ &= \sum_{i \in A_{\pi}} E_{\xi}(y_{nn2(i)} - y_{nn1(i)})^2 \end{aligned} \quad (4)$$

where A_{π} denotes the set of cases requiring imputation. The variance estimator is derived by replacing the expectation over the model in (4) with the observed values,

$$Var_{\xi}^*(Y^*) = \sum_{i \in A_{\pi}} (y_{nn2(i)} - y_{nn1(i)})^2 \quad (5)$$

Under these simple conditions, the interpretation of (5) is straightforward. The variance under the model in estimating the unobserved y_i by $y_i^* = y_{nn1(i)}$ is approximated by the squared differences of first and second nearest neighbors.

Expression of a replication method through replicate weights facilitates subsequent analysis. For the sake of generality let, Y^* , be expressed as a weighted sum,

$$Y^* = \sum_i w_{i0} y_i^* \quad (6)$$

Suppose a replication-based estimator of variance can be written in the following form,

$$Var^*(Y^*) = \sum_{r=1}^n b_r (Y_r^* - Y^*)^2, \quad (7)$$

where the b_r , $r=1, \dots, n$, are an appropriate set of coefficients independent of the choice of characteristic Y , and for replicate estimates,

$$Y_r^* = \sum_i w_{ir} y_i^*, \quad (8)$$

is defined for replicate r , on the basis of a set of replicate weights, w_{ir} , where n replicate weights are assigned to each i .

Estimator (5) can be approximated through a replicate weighting of the form (6) - (8). The Dress Rehearsal implementation used $n = 100$ and $b_r = 1$, $r = 1, \dots, 100$. For each $i \in A_{nr}$, a second record based on the second nearest neighbor, with $x = x_i$, and $y^* = y_{nn2(i)}$, can be incorporated into the data set with $w_{i0} = 0$. When the number of imputed cases is less than n , each imputed case, i , may be assigned to a unique r . Setting

$$\begin{aligned} w_{ir'} &= 1 \text{ for } y_i^* \text{ record, } r' \neq r, 0 \leq r' \leq n \\ &= 0 \text{ for } nn2 \text{ record, } r' \neq r, 0 \leq r' \leq n \\ w_{ir} &= 0 \text{ for } y_i^* \text{ record,} \\ &= 1 \text{ for } nn2 \text{ record,} \end{aligned} \quad (9)$$

exactly implements (5).

For more than 100 imputations, the cases can be serially assigned to $r = 1, \dots, 100$, losing some available precision but nonetheless producing a reasonable variance estimator with effective degrees of freedom approaching 100.

3.2 Donor Reuse The approach of 3.1 must be modified to account for use of donors more than once, which was to be required for NRFU sampling in tracts with mail response rates over 80% and for UAA sampling in general. If imputed cases i and i' share the same nearest neighbor, that is, $nn1(i) = nn1(i')$, then the expected value of the cross product is

$$\begin{aligned} E_{\xi}((y_{nn1(i)} - y_i)(y_{nn1(i')} - y_{i'})) \\ = Var_{\xi}(y_{nn1(i)} | x_{nn1(i)}) \end{aligned} \quad (10)$$

using (3). In other words, reuse of a donor contributes additional covariance affecting the variance of the estimated sum, Y^* . Estimator (5) does not incorporate the effect of this covariance.

In addition to the first 100 replicates as described, the full variance estimator incorporated two more sets of 100 replicates each that, when used jointly in (8) and (7), represented the effect of (10). To understand the general

case, it is helpful to consider first a special case. For each donor, k , let $nn1^{-1}(k)$ denote the set of imputed cases with donor k . Donor k is associated with a replicate r in 1-100, and (9) is implemented for this assignment of r for each imputed case in $nn1^{-1}(k)$.

Special Case: Suppose that, for each donor, k , whenever $nn1^{-1}(k)$ comprises $c_k > 1$ elements, each imputed case in $nn1^{-1}(k)$ has a different assigned second nearest neighbor. For this special case,

$$\begin{aligned} E_{\xi}((\sum_{i \in nn1^{-1}(k)} (y_i - y_i^*))^2) \\ = \sum_{i \in nn1^{-1}(k)} E_{\xi}(y_i - y_{nn1(i)})^2 \\ + 2 c_k (c_k - 1) Var_{\xi}(y_k | x_k) \\ = E_{\xi}((\sum_{i \in nn1^{-1}(k)} (y_{nn2(i)} - y_{nn1(i)}))^2) \end{aligned} \quad (11)$$

since for

$$\begin{aligned} E_{\xi}((y_{nn1(i)} - y_i)(y_{nn1(i')} - y_{i'})) \\ = Var_{\xi}(y_k | x_k) \end{aligned} \quad (12)$$

Thus, replicates 1-100 provide a consistent variance estimate.

When two imputed cases share both the first and second nearest neighbors, (12) no longer holds; in fact,

$$\begin{aligned} E_{\xi}((y_{nn1(i)} - y_{nn2(i)})(y_{nn1(i')} - y_{nn2(i')})) \\ = 2 Var_{\xi}(y_k | x_k) \end{aligned} \quad (13)$$

To address the resulting variance estimate in the general case, replicates 101-200 and 201-300 are included in the calculation. For any order pair of donors, k, k' , let $nnp^{-1}(k, k')$ be the set of imputations, i , with $nn1(i) = k$, and $nn2(i) = k'$, and let $c_{k,k'}$ be the number of imputations associated with the pair. For $c_{k,k'} > 1$, assign the pair to a replicate $101 \leq r \leq 200$. As in (9), for $101 \leq r' \leq 200$

$$\begin{aligned}
w_{r'} &= 1 \text{ for } y_k^* \text{ record, } r' \neq r \\
&= 0 \text{ for } nn2 \text{ record, } r' \neq r \\
w_r &= 0 \text{ for } y_k^* \text{ record,} \\
&= 1 \text{ for } nn2 \text{ record,}
\end{aligned} \tag{14}$$

These replicates are used with $b_r = -1/2$ in (7). These replicates correct effect of overestimation of the variance from the cross-product terms in (13). Unfortunately, they also subtract too much from the diagonal. To compensate, a third series of replicates $201 \leq r \leq 300$ with $b_r = 1/2$ is constructed similarly for imputed cases with $c_{k,k'} > 1$. Each imputed observation is assigned to a distinct replicate, r , for which (14) is again applied.

The presence of negative terms in the variance estimate risks negative variance estimates to a generally small degree, and it appears to increase the variance of the variance estimate compared to some alternatives. When the conditions of the Rao and Shao variance approach are met, earlier simulation of the variance estimator showed higher variance than the Rao-Shao variance formula.

The less obvious advantage to this approach is that it is directed to estimating the variance of subdomains as well as the domain total. The issue of variance estimation for subdomains was previously raised (Fay 1996) with respect to multiple imputation.

Previously cited work on this approach covers additional types of applications, including to sample surveys with negligible sampling fractions and those where the finite population correction was important. Extensions to other replication methods besides the jackknife remains an open question.

* This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. Research results and conclusions expressed are those of the author and do not necessarily indicate concurrence by the Census Bureau. It is released to inform interested parties of current research and to encourage discussion.

The author would like to thank Aref Dajani and Cary Isaki for helpful comments and Mary Ann Cochran for editorial assistance.

References

- Bankier, M., Houle, A.-M., Luc, M., and Newcombe, P. (1997), "1996 Canadian Census Demographic Variables Imputation," *1997 Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA, pp. 389-394.
- Bankier, M., Luc, M., Nadeau, C., and Newcombe, P. (1996), "Imputing Numeric and Qualitative Variables Simultaneously," *1996 Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA, pp. 90-99.
- Coder, J. (1978), "Income Data Collection and Processing for the March Income Supplement to the Current Population Survey," *Proceedings of the Data Processing Workshop: Survey of Income and Program Participation*, U.S. Department of Health, Education, and Welfare, Washington, DC.
- College, M.J., Johnson, J.H., Paré, R., and Sande, I.G. (1978), "Large Scale Imputation of Survey Data," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, DC, pp. 431-435.
- David M., Little, R.J.A., Samuël, M.E., and Triest, R.K. (1986), "Alternative Methods for CPS Income Imputation," *Journal of the American Statistical Association*, 81, 29-41.
- Dick, P. (1995), "Modelling Net Undercoverage in the 1991 Canadian Census," *Survey Methodology*, 21, 45-54.
- Farber, J. E., Fay, R.E., and Schindler, E.L. (1998), "The Statistical Methodology of Census 2000," unpublished manuscript.
- Farber, J. E. and Griffin, R. (1998), "A Comparison of Alternative Estimation Methodologies for Census 2000," *1998 Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA, pp. 629-634.
- Fay R.E. (1996), "Alternative Paradigms for the Analysis of Imputed Survey Data," *Journal of the American Statistical Association*, 91, 490-498.
- Fay, R. E. and Farber, J. (1999), "The Census 2000 Dress Rehearsal: Methodological Basis for Nonresponse Followup Estimation," to be presented at the Federal Committee on Statistical Methodology Research Conference, Arlington, VA, Nov. 15-17, 1999.
- Fay, R. E. and Town, M. K. (1996), "Variance Estimation for the 1995 Census Test: Methodology and Findings," in *Proceedings of the 1996 Annual Research Conference*, U.S. Bureau of the Census, pp. 761-781.

- _____. (1998), "Variance Estimation for the 1998 Census Dress Rehearsal," *1998 Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA, pp. 605-610.
- Ford, Barry L. (1983), "An Overview of Hot-Deck Procedures," Madow, W.G., Olkin, I., and Rubin, D.B. (eds.), *Incomplete Data in Sample Surveys, Vol. 2*, National Research Council, Panel on Incomplete Data, Academic Press, New York, pp. 185-207.
- Germain, M.-F. and Julien, C. (1993), "Results of the 1991 Census Coverage Error Measurement Program," *Proceedings of the Seventh Annual Research Conference*, U.S. Bureau of the Census, pp. 55-70.
- Lee, H., Rancourt, E., and Särndal, C.E. (1994), "Experiments with Variance Estimation from Survey Data with Imputed Values," *Journal of Official Statistics*, 10, 231-243.
- Navarro, A., Treat, J., and Mulry, M. (1996), "Nonresponse Followup: Unit vs. Block Sampling," *1996 Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA, pp. 551-556.
- Rao, J.N.K. and Shao, J. (1992), "Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation," *Biometrika*, 79, 811-822.
- Rizvi, M.H. (1983), "An Empirical Investigation of Some Item Nonresponse Adjustment Procedures," in Madow, W.G., Nisselson, H. and Olkin, I. (eds.), *Incomplete Data in Sample Surveys, Vol. 1*, National Research Council, Panel on Incomplete Data, Academic Press, New York, pp. 299-366.
- Robinson, J.G., Ahmed, B., Das Gupta, P., and Woodrow, K.A., "Estimation of Population Coverage in the 1990 United States Census Based on Demographic Analysis," *Journal of the American Statistical Association*, 88, 1061-1071.
- Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley.
- Sande, G.T. (1983), "Replacement for a Ten-Minute Gap," in Madow, W.G. and Olkin, I., *Incomplete Data in Sample Surveys, Vol. 2*, National Research Council, Panel on Incomplete Data, Academic Press, New York, pp. 337-338.
- Sande, I.G. (1983), "Hot-Deck Imputation Procedures," in Madow, W.G. and Olkin, I., *Incomplete Data in Sample Surveys, Vol. 2*, National Research Council, Panel on Incomplete Data, Academic Press, New York, pp. 339-349.
- Schaefer, J.L. (1995), "Model-Based Imputation of Census Short-Form Items," in *Proceedings of the 1995 Annual Research Conference*, U.S. Bureau of the Census, pp. 267-299.
- Steel, P. and Fay, R.E. (1995), "Variance Estimation for Finite Populations with Imputed Data," *1995 Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA, pp. 374-379.
- Thompson, J. H. and Fay, R. E. (1998), "Census 2000: The Statistical Issues," *1998 Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA, pp. 101-110.
- Town, M.K., and Fay, R.E. (1995), "Properties of Variance Estimators for the 1995 Census Test," *1995 Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA, pp. 724-729.
- Tsay, J.H., Isaki, C.T., and Fuller, W.A. (1996), "A Block Based Nonresponse Followup Survey Design," *1996 Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA, pp. 557-562.
- U.S. Census Bureau (1999a), "Contamination of Initial Phase Data Collected in ICM Block Clusters," by Sam Hawala, Census 2000 Dress Rehearsal Evaluation Memorandum C2, Executive Summary available from <http://www.census.gov/census2000/evaluations/pdf/sumc2.pdf>.
- _____. (1999b), "Specifications for Nonresponse Followup and Undeliverable-as-Addressed Vacant Estimation in the Census 2000 Dress Rehearsal," DSSD Census 2000 Dress Rehearsal Memorandum Series # A-40, from Donna L. Kostanich to Dennis W. Stoudt.
- Wright, T. (1999), "A one-number census: some related history," *Science*, 283, pp. 491-492.
- Zanutto, E. and Zaslavsky, A.M. (1996), "Estimating a Population Roster from an Incomplete Census, Using Mailback Questionnaires, Administrative Records, and Sampled Nonresponse Followup," *1996 Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA, pp. 538-543.

ERRONEOUSLY ENUMERATED PEOPLE IN THE CENSUS 2000 DRESS REHEARSAL

Roxanne Feldpausch and Danny R. Childers¹
Roxanne Feldpausch, Bureau of the Census, Washington, DC 20233

Key words: Integrated Coverage Measurement, Dual System Estimation

I. Introduction

The Census 2000 Dress Rehearsal was conducted in three sites: Sacramento, California; Menominee county, Wisconsin (mainly Menominee Indian Reservation); and Columbia and its surrounding areas in South Carolina. In Sacramento and Menominee, the census was conducted in two parts. First, the initial phase enumerated the people. Then the Integrated Coverage Measurement (ICM) estimated the people missed in the initial phase. The combination was the census count. In South Carolina a traditional census was conducted. After the census, there was a Post-Enumeration Survey to estimate the coverage of the census. In this paper both the initial phase and the traditional census are referred to as a census and both of the coverage surveys are referred to as the ICM.

The ICM estimated, among other things, how many people enumerated in the census were enumerated in error. The number of erroneous enumerations is one of the inputs into the dual system estimator, which is a factor used to determine the final census count (Schindler, 1999). In this paper, we look at various factors which may be related to a person's probability of being erroneously enumerated.

To determine the number of erroneous enumerations in the census, the E-sample people (the people captured in the census) were matched to the people captured in the ICM. After the computer and clerical matching phase, the E-sample people were classified as matched, not matched, or possibly matched. Those who matched were considered correctly enumerated. The nonmatched E-sample people were followed up to determine if they were correctly or erroneously enumerated in the block cluster according to census residence rules. If the follow-up interview could not determine the person to be correctly or erroneously enumerated, the enumeration status for the E-sample person was unresolved. Those people

with unresolved enumeration status had their erroneous enumeration probabilities imputed.

Section II discusses the methods used to analyze the data. Section III examines the erroneous enumeration rates of various subgroups. Sections IV examines various types of erroneous enumerations. Section V summarizes the findings.

II. Methodology

For analysis purposes we broke the South Carolina site up into three distinct areas: the city of Columbia, referred to as Columbia; the remaining counties which were mail-out/mail-back, referred to as Other SC; and the counties that were update/leave, referred to as Rural SC. For mail-out/mail-back areas, the mailing list was obtained from the US Postal Service, 1990 Census, local, tribal, and other potential supplementary address sources. A census questionnaire was mailed to the addresses and if occupied, the residents were to mail back a completed form. For update/leave areas the enumeration procedures were different. The address list was obtained by Census Bureau employees who created a listing of addresses before the census. The enumerators updated the census address list and left a questionnaire for the household to complete and mail back to the Census Bureau.

In these five different areas, we examined the erroneous enumeration rates of people in various subsets of the population. We estimated the erroneous enumeration rate by the number of erroneous enumerations divided by the total number of people in the E sample. For the erroneous enumeration rates, we calculated the standard errors using the simple Jackknife method. The simple Jackknife should yield standard errors similar to those produced with the stratified Jackknife which was used in the Dress Rehearsal. These standard errors were computed using the statistical package VPLX. The internet site www.census.gov/sdms/www/vwelcome.html has more information on VPLX.

Once we computed standard errors, we used a two-

¹ Roxanne Feldpausch and Danny Childers are mathematical statisticians in the Decennial Statistical Studies Division of the US Census Bureau. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. It is released to inform interested parties of research and to encourage discussion.

tailed t-test to determine which differences were significant. When there were multiple comparisons, we used the Bonferroni adjustment to determine which comparisons were significant (Games, 1971). All tests were conducted at a 0.10 significance level. The tests were conducted under the assumption that the observations were independent of each other. The analyses were conducted on the final data, after weighting and imputation. All numbers in this paper are weighted and the probability of erroneous enumerations for people with unresolved enumeration status are included in the percent of erroneous enumerations.

III. Percentage Erroneous Enumeration

For the 1998 Dress Rehearsal, the rate of erroneous enumerations varied across the areas from 7.7 percent in Columbia to 16.5 percent in Other SC. In the 1990 Census, the erroneous enumeration rate for the nation was 4.6 percent (Griffin and Moriarity, 1992). Table 1 shows the percentage of erroneous enumerations out of the total number E sample people for each site. It also gives the number of erroneous enumerations in each site. To compare the percent of erroneous enumerations among the different areas, we used the Bonferroni adjustment. The adjusted alpha is 0.010 and the criterion t-value is 2.560. Other SC had a significantly higher erroneous enumeration rate than Sacramento ($t=3.063$), Columbia ($t=4.202$) and Menominee ($t=2.762$). The remaining areas did not differ in their erroneous enumeration rates.

Table 1: Percentage (Standard Error) and Number of Erroneous Enumerations by Site

Site	Percent EE	Number of EE
Sacramento	10.5 (0.71)	38,878
Columbia	7.7 (1.02)	6,438
Rural SC	10.3 (1.84)	14,858
Other SC	16.5 (1.82)	58,837
Menominee	9.8 (1.62)	321

One group of characteristics that we examined was the poststrata variables for the dual system estimator. These variables are: tenure, sex, age and race. Tenure has been shown to be related to erroneous enumeration rates. In the 1990 Census, renters were more likely to be enumerated in error than owners (Griffin and Moriarity, 1992). In Sacramento ($t=4.818$), Rural SC ($t=1.691$), and Other SC ($t=2.386$) this held true. For these three areas, owners had significantly lower erroneous enumeration rates than renters. In Columbia ($t=0.384$) and Menominee ($t=1.240$) there were no differences between the erroneous enumeration rates of

renter and owners. See Table 2 for the percentage of erroneous enumerations by renter and owner.

Table 2: Percentage of Erroneous Enumerations (Standard Error) by Tenure

Site	Owner	Renter
Sacramento	7.8 (0.56)	13.6 (1.20)
Columbia	7.2 (0.63)	8.1 (2.21)
Rural SC	9.2 (1.50)	15.8 (4.74)
Other SC	13.2 (1.24)	24.6 (4.82)
Menominee	11.5 (1.88)	6.6 (3.11)

There were no significant difference in erroneous enumeration rates between males and females (see Table 3) across all of the Dress Rehearsal areas. This differs from 1990 results for the nation as a whole in which males had a higher erroneous enumeration rate than females (Moriarity, 1993).

Table 3: Percentage of Erroneous Enumerations (Standard Error) by Sex

Site	Male	Female
Sacramento	10.8 (0.71)	10.3 (0.74)
Columbia	7.7 (1.07)	7.8 (1.03)
Rural SC	10.7 (1.76)	9.9 (1.98)
Other SC	16.0 (1.61)	17.0 (2.06)
Menominee	9.8 (1.78)	9.7 (1.75)

Age, on the other hand, was related to the erroneous enumeration rate. In the Dress Rehearsal as in the 1990 Census, there were four age poststratification categories: 0-17, 18-29, 30-49, and 50 and over (see Table 4). The Bonferroni adjustment for the four groups (six comparisons) produced an adjusted alpha of 0.017 and the criterion t-value of 2.378.

Across the areas, the 18-29 age group tended to have higher rates of erroneous enumeration than other age groups. In Sacramento, those 18-29 had a higher erroneous enumeration rate than the 30-49 year olds ($t=2.895$) and those 50 and over ($t=5.466$). People, in Sacramento, 50 and over had a lower erroneous enumeration rate than 0-17 year olds ($t=3.899$) and 30-49 year olds ($t=4.895$). In Columbia, 18-29 year olds also had a higher erroneous enumeration rate than both the 30-49 year olds (2.671) and those 50 and over ($t=2.627$). In Rural SC, 18-29 year olds had a higher rate of erroneous enumeration than the 0-17 year olds ($t=2.590$) and the 30-49 year olds ($t=2.474$). In Other SC, the 18-29 year olds had a higher erroneous enumeration rate than the 39-49 age group ($t=2.867$). In Menominee we found a different pattern. Those 50 and older had a higher erroneous enumeration rate than 18-29 year olds ($t=2.438$). Across the Dress Rehearsal

areas, we found no other age categories to be significantly different.

Table 4: Percentage of Erroneous Enumerations (Standard Error) by Age

Site	0-17	18-29	30-49	50+
Sacramento	11.3 (1.06)	12.8 (1.00)	10.7 (0.64)	8.0 (0.63)
Columbia	7.0 (2.31)	9.6 (0.90)	7.6 (1.03)	7.0 (0.58)
Rural SC	9.6 (2.08)	13.8 (2.55)	9.2 (1.64)	10.4 (2.24)
Other SC	18.2 (3.08)	19.7 (2.46)	14.7 (1.44)	15.1 (1.54)
Menominee	8.9 (2.43)	7.4 (3.16)	8.0 (1.84)	13.4 (1.75)

We used different race categories in the various sites reflecting their differing racial makeups. People who marked more than one racial category were assigned to the largest nonwhite category that they marked based on 1990 Census numbers. People who did not mark any racial category were assigned to the non-Hispanic white category. Race/origin groups with less than one percent of the site's 1990 population total were collapsed into the largest nonwhite race according to 1990 data. See Schindler (1999) for a more complete explanation of the racial categories.

Table 5: Percentage of Erroneous Enumerations (Standard Error) in Sacramento by Race

Non-Hispanic White	Non-Hispanic Black	Non-Hispanic Asian	Hispanic
9.0 (0.55)	15.0 (1.32)	10.4 (1.60)	10.6 (0.91)

In Sacramento there were four race categories: non-Hispanic white, non-Hispanic black, non-Hispanic Asian, and Hispanic. All other races were collapsed with Hispanics for estimation purposes. The Bonferroni adjustment for the four groups (six comparisons) produced an adjusted alpha of 0.017 and the criterion t-value of 2.378. The erroneous enumeration rates for the various racial categories in Sacramento ranged from 9.0 percent for non-Hispanic whites to 15.0 percent for non-Hispanic blacks as shown in Table 5. Non-Hispanic blacks had significantly higher erroneous enumeration rate than non-Hispanic whites ($t=4.868$), non-Hispanic Asians ($t=3.450$), and Hispanics (4.052). The other racial categories did not differ from each other in their erroneous enumeration rates.

In the South Carolina site there were two racial

categories: non-Hispanic white and black. All other race groups were collapsed with blacks for estimation purposes. In Columbia ($t=1.333$), Rural SC ($t=1.010$) and Other SC ($t=0.708$) there were no significant differences between the erroneous enumeration rates of non-Hispanic whites and blacks. See Table 6 for the percentage of erroneous enumerations in each of the three South Carolina areas.

Table 6: Percentage of Erroneous Enumerations (Standard Error) in South Carolina by Race

Site	Non-Hispanic White	Black
Columbia	6.7 (0.62)	8.8 (1.80)
Rural SC	8.9 (1.42)	13.0 (4.12)
Other SC	15.4 (1.53)	17.9 (3.41)

In Menominee there were two racial categories: non-Hispanic white and American Indian. All other race groups were collapsed with American Indians for estimation purposes. As seen in Table 7, American Indians had a 7.4 percent erroneous enumeration rate, while nearly 20 percent of the non-Hispanic white people were erroneously enumerated. The American Indians had a significantly lower erroneous enumeration rate than non-Hispanic whites ($t=6.346$).

Table 7: Percentage of Erroneous Enumerations (Standard Error) in Menominee by Race

Non-Hispanic White	American Indian
19.8 (1.33)	7.4 (1.47)

When analyzing the erroneous enumeration rates, we considered factors other than the poststrata variables used in the dual system estimator. We also looked at variables related to form such as form length and return type. There were two different form lengths in the Dress Rehearsal: short and long. Approximately 17% of the housing units received a long form which asked for more detailed information about the housing unit and the people living there. The percentage of erroneous enumerations by form type are shown below in Table 8. There was no significant difference in the erroneous enumerations rate of those people who filled out short forms and those who filled out long forms. This is consistent with the 1990 Census results (Griffin and Moriarty, 1992).

Table 8: Percentage of Erroneous Enumerations (Standard Error) by Form Length

Site	Short	Long
Sacramento	10.4 (0.71)	11.0 (1.05)
Columbia	7.6 (1.05)	8.7 (1.19)

Rural SC	9.8 (1.43)	12.5 (4.43)
Other SC	16.4 (1.89)	17.3 (1.92)
Menominee	9.4 (1.64)	12.9 (4.92)

During the Dress Rehearsal, households received a questionnaire which they were supposed to complete and mail back. Those households who did not return their questionnaire were visited by an enumerator who collected the information. For those returns filled out by an enumerator, the information about the household could have come from a household member or it could have been obtained through a proxy interview. For some interviews, the enumerator failed to indicate whether or not the respondent was a proxy. There were four groups, so the adjusted alpha is 0.017 and the criterion t is 2.370. We analyzed Menominee separately due to the small sample size.

We found that mail returns had lower erroneous enumeration rates than both proxy and non-proxy enumerator filled returns. We found that non-proxy enumerator filled returns had lower erroneous enumeration rates than proxy enumerator filled returns. We also found that enumerator filled returns where the proxy information was missing had higher erroneous enumeration rates than mail returns and non-proxy enumerator returns. See Table 9 for the percentage of erroneous enumerations for these variables.

In Sacramento ($t=10.411$), Columbia ($t=3.622$), Rural SC ($t=3.353$) and Other SC ($t=2.406$) the mail returns had significantly lower erroneous enumeration rates than non-proxy enumerator filled returns.

In Sacramento ($t=11.841$), Columbia ($t=6.281$), Rural SC ($t=3.315$) and Other SC ($t=4.657$) the mail returns had significantly lower erroneous enumeration rates than proxy enumerator filled returns.

In Sacramento ($t=8.979$), Columbia ($t=5.760$), Rural SC ($t=2.678$) and Other SC ($t=3.408$) non-proxy enumerator filled returns had significantly lower erroneous enumeration rates than proxy enumerator filled returns.

In Sacramento ($t=4.801$), Columbia ($t=3.303$) and Other SC ($t=2.882$) mail returns had significantly lower erroneous enumeration rates than those enumerator returns where the proxy information was missing. In Rural SC ($t=1.570$) there was no difference.

In Sacramento ($t=2.470$), Columbia ($t=2.760$) and Other SC ($t=2.419$) non-proxy enumerator filled returns had lower erroneous enumeration rates than those enumerator filled returns where the proxy information was missing. In Rural SC ($t=0.977$) there was no difference.

In Sacramento ($t=3.700$) those enumerator filled returns where the proxy information was missing had a lower erroneous enumeration rate than proxy

enumerator filled returns.

For Menominee, we only considered mail returns versus enumerator filled returns. Mail returns had 8.8 (1.74) percent erroneous enumerations while enumerator filled returns had 10.3 (3.53) percent erroneous enumerations. There was no difference ($t=0.231$) between the erroneous enumeration rates of mail returns and those of enumerator filled returns.

Table 9: Percentage of Erroneous Enumerations (Standard Error) by Return Type

Site	Mail	Enumerator		
		No Proxy	Proxy	Missing
Sacramento	6.5 (0.65)	14.1 (0.96)	38.5 (2.77)	22.3 (3.32)
Columbia	6.0 (0.85)	9.0 (1.38)	28.4 (3.68)	22.1 (4.92)
Rural SC	7.0 (1.53)	13.1 (2.56)	33.4 (8.10)	23.5 (10.58)
Other SC	14.1 (1.51)	19.8 (3.14)	34.2 (4.49)	42.4 (10.11)

Next, we looked at the number of people in a household to see if that was related to the erroneous enumeration rate of a person in the household. We compared three groups of households: 1 person, 2-5 people and 6 or more people (see Table 11).

Table 11: Percentage of Erroneous Enumerations (Standard Error) by Number of People in the Household

Site	1 person	2-5 people	6+ people
Sacramento	12.4 (0.88)	10.1 (0.73)	10.9 (1.43)
Columbia	9.2 (0.67)	7.5 (1.21)	7.0 (1.61)
Rural SC	13.9 (2.48)	10.0 (1.76)	8.4 (6.68)
Other SC	18.9 (2.24)	16.3 (1.80)	14.4 (4.74)
Menominee	11.5 (3.89)	10.0 (1.60)	8.2 (4.16)

The Bonferroni adjustment for three comparisons produced an adjusted alpha of 0.035 and a criterion t of 2.114. In Sacramento ($t=2.952$) we found that people in single person households had a significantly higher erroneous enumeration rate than those people in two to five people households. There were no other differences. This is different from 1990 where it was found that erroneous enumeration rates increased as household size increased (Griffin and Moriarity, 1992).

For those people who did not answer all of the

questions on their Census form, the missing values were imputed (see Table 12). It appears that complete data had less error. In Sacramento ($t=13.960$), Columbia (5.093), Rural SC ($t=3.072$), and Other SC (3.800) those people with at least one item imputed had significantly higher erroneous enumeration rates than those people with no imputed items. In Menominee ($t=0.208$) there was no difference.

Table 12: Percentage of Erroneous Enumerations (Standard Error) by

Site	No Imputations	Some Imputations
Sacramento	6.0 (0.57)	16.1 (0.98)
Columbia	5.5 (0.85)	11.1 (1.62)
Rural SC	7.3 (1.55)	13.6 (2.56)
Other SC	13.3 (1.45)	21.7 (2.87)
Menominee	10.1 (1.67)	9.4 (2.89)

III. IV. Types of Erroneous Enumeration

There are many reasons why a person could be an erroneous enumeration: geocoding errors, fictitious people, duplicate records, other counting errors, insufficient information for matching and unresolved cases.

Geocoding errors occurred when the census placed a housing unit in the wrong block. All of the people in that housing unit were then considered geocoding errors. A fictitious person was another type of erroneous enumeration. A census person could be found to be fictitious if follow-up (after the ICM) determined that the person did not refer to a real person in that block. Duplicates occurred when a person had two or more census records. These additional records were duplicates. Other counting errors occurred when a person was counted in error in a block cluster in the census. The ICM then determined that according to census residency rules the person should have been counted elsewhere. For example, a college student counted in the wrong place or a family with two homes is an other counting error.

People with insufficient information for matching were treated as erroneous enumerations. To have sufficient information for matching, a person had to have had a name and at least one other characteristic provided. People without these two pieces of information were considered to have insufficient information for matching.

An unresolved case occurred when there was not enough information to determine if the person was correctly or erroneously enumerated. A case could have been unresolved because not enough information was collected during the ICM to determine whether or not

the person was correctly enumerated during the census. Another example of an unresolved case is a person who was a match, but had an unresolved residency status. Finally, a person who was a possible match, but did not have enough information to positively determine their match status was unresolved. The unresolved cases had their erroneous enumeration probability imputed using a proportion of erroneous enumerations from those people resolved during follow-up. For more information on these categories see Childers (1998).

In Sacramento, geocoding errors and insufficient information for matching were the major causes of erroneous enumeration accounting for about 64 percent of the erroneous enumerations as seen in Table 13.

Table 13: Percentage of Different Types of Erroneous Enumerations (Standard Error) in Sacramento

Type of Error	Percentage
Geocoding error	28.2 (5.08)
Fictitious	10.3 (1.50)
Duplicates	10.3 (1.20)
Other Counting Error	9.5 (1.03)
Insufficient Information	35.4 (2.73)
Unresolved	6.3 (0.66)

In the various South Carolina areas the make-up of the erroneous enumerations varied. In Columbia (see Table 14), insufficient information for matching accounted for approximately 32 percent of the erroneous enumerations. Geocoding errors accounted for about 26 percent of the erroneous enumerations.

Table 14: Percentage of Different Types of Erroneous Enumerations (Standard Error) in Columbia

Type of Error	Percentage
Geocoding Error	26.2 (7.03)
Fictitious	8.1 (1.52)
Duplicates	14.0 (3.36)
Other Counting Error	13.4 (2.01)
Insufficient Information	31.6 (3.61)
Unresolved	6.7 (1.32)

In Rural SC, geocoding errors accounted for about 40 percent of the erroneous enumerations. Duplicates accounted for about 23 percent and other counting errors accounted for about 16 percent of erroneous enumerations. See Table 15 for the different types of erroneous enumerations for Rural SC.

Table 15: Percentage of Different Types of Erroneous Enumerations (Standard Error) in Rural SC

Type of Error	Percentage
Geocoding Error	40.1 (11.97)
Fictitious	9.3 (4.08)

Duplicates	22.9 (6.04)
Other Counting Error	15.7 (4.25)
Insufficient Information	10.2 (2.71)
Unresolved	1.9 (0.61)

In Other SC, geocoding errors accounted for about 64% of the erroneous enumerations (see Table 16). Other SC was update/leave. The different method of enumeration may be the reason for the relatively high percentage of geocoding errors. People with insufficient information for matching accounted for about 12 percent of the erroneous enumerations in Other SC.

Table 16: Percentage of Different Types of Erroneous Enumerations (Standard Error) in Other SC

Type of Error	Percentage
Geocoding Error	63.5 (6.24)
Fictitious	5.7 (2.42)
Duplicates	7.9 (1.67)
Other Counting Error	8.0 (1.46)
Insufficient Information	11.6 (1.75)
Unresolved	3.4 (1.45)

In Menominee, over half (50.8 percent) the erroneous enumerations were due to duplicate records for the same person (see Table 17). Other counting errors contributed to about 32 percent of the erroneous enumerations.

Table 17: Percentage of Different Types of Erroneous Enumerations (Standard Error) in Menominee

Type of Error	Percentage
Geocoding Error	1.3 (1.38)
Fictitious	1.3 (1.00)
Duplicates	50.8 (8.82)
Other Counting Error	31.5 (10.40)
Insufficient Information	13.5 (6.32)
Unresolved	1.6 (0.85)

IV. Conclusions

The percentage of erroneous enumerations varied across the sites ranging from 7.7 percent in Columbia to 16.5 percent in Other SC. We determined erroneous enumeration rates for various subgroups of the population. We looked at the poststrata categories, form characteristics and other variables.

In general, we found that renters had higher erroneous enumeration rates than owner and 18-29 year olds had higher erroneous enumeration rates than other age groups. We found no differences in the erroneous enumeration rates between the sexes.

Form length was not related to erroneous enumeration rates. However, return type was related to erroneous enumeration rates. Mail returns had lower erroneous enumeration rates than enumerator filled returns. For enumerator filled returns, non-proxy responses had lower erroneous enumeration rates than proxy responses. We also found that those people with some variables imputed had higher erroneous enumeration rates than those people with no variables imputed. Proxy responses and imputed values are an indicator of poor data quality. Although these results are only representative of the Dress Rehearsal sites, they do show that the quality of the data is related to the erroneous enumeration rate. It is difficult to draw conclusions about the 2000 Census from these data because the census methods are not the same as in the Dress Rehearsal.

V. References

- Childers, Danny (1998), "The Design of the Census 2000 Dress Rehearsal Integrated Coverage Measurement," *DSSD Census 2000 Dress Rehearsal Memorandum Series, Chapter F-DT-2*.
- Games, Paul (1971) "Multiple Comparisons of Means," *American Educational Research Journal*, 8, 3, 531-565.
- Griffin, Deborah and Moriarity, Christopher (1992) "Characteristics of Census Errors," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 512-517.
- Moriarity, Christopher (1993), "Characteristics of Census Error - Additional Results," *1990 Decennial Census Preliminary Research and Evaluation Memorandum Number 240*.
- Schindler, Eric (1999) "Iterative Proportional Fitting in the Census 2000 Dress Rehearsal," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, to appear.

MODELING CENSUS AND INTEGRATED COVERAGE MEASUREMENT PHASE MISSES IN THE CENSUS 2000 DRESS REHEARSAL

Michael Beaghen¹, Bureau of the Census
Washington, D.C., 20233

KEY WORDS: logistic regression, E-sample, P-sample.

I. Introduction

The Census 2000 Dress Rehearsal was conducted at three sites: Sacramento, California, Menominee, Wisconsin, and Columbia South Carolina and its surrounding counties. In this paper Columbia City is treated as a site separate from its surrounding counties, and Menominee is not included due to its small size. The methodology differed in the Columbia sites from the Sacramento site. In the Columbia sites a traditional census and a subsequent Post Enumeration Survey (PES) were performed. In Sacramento the Integrated Coverage Measurement (ICM) was performed which consisted of two phases: an initial phase akin to a traditional census and a secondary survey akin to a PES. For the purposes of this analysis the methodologies are treated the same. The primary survey shall be referred to as the census and the secondary survey shall be referred to as the ICM. The census enumerated the entire sites. After the census was completed, the ICM performed an independent reenumeration conducted only in selected block clusters within the sites. In Sacramento a single estimate was produced based on the census and the ICM. In South Carolina an estimate based on the census and an estimate based on the PES was produced.

For the ICM sample the block clusters were the primary sampling units. They were drawn from twelve sampling strata. Those selected block clusters with 80 or more housing units were subsampled. The selected block clusters after subsampling comprised the ICM sample.

The census enumerations within the ICM sample block clusters define the E-sample. The people enumerated in the ICM in the sample block clusters define the P-sample. A matching operation linked E-sample people with P-sample people. A linked pair is called a match. An E-sample enumeration that was not linked to a P-sample

person was an E-sample non-match. A follow-up interview determined whether the person existed in the ICM sample. People found to exist are called confirmed non-matches. People found not to exist in the sample are called erroneous enumerations. Erroneous enumerations will be ignored in this analysis. Confirmed E-sample non-matches represent P-sample misses or failures to capture.

A P-sample person that does not match to an E-sample enumeration is called a P-sample non-match. A follow-up interview determined whether the person existed in the ICM sample. Confirmed P-sample non-matches represent E-sample misses or failures to capture.

The purpose of this paper is to use logistic regression models to relate these P-sample misses and E-sample misses to demographic characteristics and housing unit characteristics. The limitation of univariate descriptive statistics is that they do not address the question of the relationship of one variable in the context of other variables. A regression type model avoids this limitation. Since the response is binary, that is, a person is either captured or missed, logistic regression is an obvious method.

This study is observational rather than experimental. The characteristics used as regressors in the model are not controlled by the researcher but rather are random variables. Consequently the modeling is not predictive but descriptive and the hypothesis tests used to determine which variables to include in the model are not strictly correct. They are to be understood as guidelines in model building.

The paper is organized as follows. First, the variables are laid out and described. Then I build two models. To model the P-sample misses, I model the E-sample people who were matched to ICM people against those who were confirmed non-matches. The erroneous

¹Michael Beaghen is a mathematical statistician in the Decennial Statistical Studies Division of the US Census Bureau. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. The research results and conclusion are expressed are those of the author and do not necessarily indicate concurrence by the Census Bureau. It is released to inform interested parties of current research and to encourage discussion.

enumerations were discarded from the analysis. To model the E-sample misses, I model the P-sample people who were matched against those who were confirmed non-matches.

II. The Variables

In both analyses the response variable has two outcomes: a person is a match or a person is a non-match that is confirmed to exist.

The values of the regressor variables in both data sets were formatted such that they would have the same categories, allowing the models to be comparable. All of the regressor variables are dichotomized. Categorical variables indicating group membership (in italics) are represented as dichotomous indicator variables, with one category serving as a baseline to which all the other categories are compared. For example, for the variable race, white serves as the baseline. Age, the only continuous variable, was formatted as a categorical variable. The breakdown of the categorical variables into classes is as follows.

Race: *black* (black or African-American), *Asian* (Asian or Pacific Islander), *mixed* (mixed ancestry, American Indian or other); white is the reference group.

Age: *child* (age ≤ 17), *young* ($18 \leq \text{age} \leq 29$), *middle* ($30 \leq \text{age} \leq 49$); older ($50 \leq \text{age}$) is the reference group.

Tenure: *renter* (all nonowners); owner is the reference group.

Site: *Columbia*, *RuralSC*; Sacramento is the reference group.

Relationship to Respondent:

family (person was an immediate family member of respondent), *relative* (person was a non-immediate family member of the respondent), *other* (a non-relative, or relative beyond immediate family or cousins, grandparents, or aunts or uncles); the respondent serves as the reference group.

Multi-unit:

multi-unit (person lived in a multi-unit); the reference group is formed by those living in single units, which includes trailers.

III. Methods

I used SUDAAN's (Shah, Barnwell & Bieler, 1997) Proc Logistic to estimate the models. SUDAAN is designed to handle data from complex surveys such as the ICM.

I also used SAS (1989) software's Proc Logistic to estimate the models. SAS has more modeling capabilities than SUDAAN. To account for the complex survey design two measures were taken. First, the observations were weighted by the inverse of the sampling weight. This made the estimates reflect the population of the sites, not the sample. Second, the Wald test statistic was divided by a design effect since SAS estimates variances as if the sample were a simple random sample. This method gives the correct maximum likelihood estimates but the Wald test statistics are approximate at best.

The SUDAAN modeling is viewed as more correct and generated the estimates seen in the tables.

The backward, forward and stepwise methods of model selection yielded similar models, though the backward selection models were superior for both models. The cutoff for removing a variable was a p-value greater than 0.03. This choice was somewhat arbitrary. It allowed me not to include some variables that had only very weak effects.

IV. Interpretation - Modeling the P-Sample Misses

In logistic regression, the binary response is thought of as successes versus failures. In this model define a success as a person in the E-sample but not on the P-sample, i.e., a P-sample miss. Define a failure as a match. Rather than examining the parameter weights themselves it is easier to interpret the odds ratios associated with an increase of one unit for each parameter. These odds ratios are directly related to the parameter weights. Since each of the variables has a value of zero or one, depending on group membership as described in Section 2., the interpretation of the odds ratios are straightforward. The odds ratio will refer to the ratio of the odds of a response with a value of one with the odds of a response with a value of zero. For example, see the variable *renter* in Table 1. *renter* has an odds ratio of 1.48. Since a value of one indicates a person lives in a rented unit and a value of zero indicates a person lives in an owned unit, the odds ratio shows that a *renter* has a 48% greater odds of being missed by the P-sample than an owner, all other variables being held constant.

If the variable of interest is an indicator variable in a

series of indicator variables that denote membership to a group the odds ratio refers to the comparison with the baseline group (Hosmer, Lemeshow 1989). As an illustration, consider race. As shown in Section 2., the four race categories, white, black, Asian, and mixed, are indicated by three indicator variables, *black*, *Asian*, and *mixed*. Only *black* and *mixed* are in the model. The odds ratio of 1.31 associated with those of mixed ancestry implies that a person of mixed ancestry has a 31% greater odds of being found in the E-sample and missed in the P-sample than those a person of white ancestry.

Another factor to consider in assessing the relative importance of a variable is how often it has a value of one. For example, although *black* has an odds ratio a good bit higher than *male*, there are only a fraction as many people who are *black* as *male*.

To summarize the results of the model, race, tenure,

multi-unit, age, relationship to respondent, sex and blocksize all are associated with capture in the P-sample. Renters are more likely to be missed than are non-renters. Non-immediate relatives and non-relatives are more likely to be missed than the respondent or the respondent's immediate family (i.e., the variable indicating immediate family was not significant). People of black or mixed ancestry are more likely to be missed than those of white or Asian ancestry. People who live in multi-units are more likely to be missed than people who live in single units. Children and young adults are more likely to be missed than middle aged or older people. Males were slightly more likely to be missed than females.

Also illuminating are the several categorical variables that are non-significant. They are Hispanic origin and all second level interaction terms.

Table 1. Odds Ratios

Reference Variable	Model Variable	P-Sample Misses Odds Ratios	E-Sample Misses Odds Ratios
Owner	<i>Renter</i>	1.48	1.32
Non-Hispanic	<i>Hispanic</i>	ns	ns
Female	<i>Male</i>	1.09	1.05
Older	<i>Child</i>	1.58	1.21
	<i>Young</i>	1.24	1.33
	<i>Middle Age</i>	ns	ns
Respondent	<i>Relative</i>	1.87	1.36
	<i>Family</i>	ns	ns
	<i>Other</i>	1.56	1.68
White	<i>Black</i>	1.31	1.54
	<i>Asian</i>	ns	1.43
	<i>Mixed</i>	1.31	1.36
Single-Unit	<i>Multi-Unit</i>	1.59	1.73
Sacramento	<i>RuralSC</i>	ns	1.56
	<i>Columbia</i>	ns	0.65
Non-Large	<i>Large Block</i>	1.39	ns

V. Modeling the E-Sample Misses

A P-sample confirmed non-match represents a person missed by the census, that is an E-sample miss. In this second model define a success as an E-sample miss and define a failure as a match. I will compare this model, which models census misses, to the model of the previous section which modeled the P-sample misses. It is reasonable to ask if the people missed in the ICM and the Census are similar in their characteristics because both eluded similar surveys. Upon examination of the odds ratios this seems to be true to an extent, though there were some important differences in the models.

Firstly, a person in the Columbia city site was less likely to be missed by the census than a person in the Sacramento site, though a person in the rural South Carolina site had the greatest chance of all of being missed by the census. In comparison, the P-sample site was not related to the miss rate.

Secondly, P-sample people were more likely to be missed in large blocks, though this was not true for E-sample people.

Except for the three situations just described, the characteristics describing E-sample misses are similar to those describing P-sample misses. Age, relationship to respondent, race, tenure and multi-unit are similar in the nature of their association with misses. Likewise, in both surveys sex and Hispanic origin played little or no role. See Table 1. which shows the odds ratios for each variable.

VI. Conclusion

Logistic regression is a useful method to examine what variables are associated with P-sample and E-sample misses. Not surprisingly, many of the same variables associated with P-sample misses are associated with E-sample misses. They are age, race, tenure, multi-unit and relationship to the respondent. The role site played differed in the ICM and census. Hispanic origin played no role and sex played a small but similar role in either census or ICM capture. Second order interactions were not statistically significant. This suggests that had I modeled separately for Sacramento and South Carolina that I would have similar results.

This work could have implications for the poststratification. Poststrata are defined such that the probability of capture is as homogeneous as possible within each poststratum, for both P-sample capture and

E-sample capture (Wolter 1986). In the 1998 ICM poststratification was done by tenure, age, sex, race and Hispanic origin. The results here suggest different poststrata. Sex and Hispanic origin were of little use in discriminating capture probabilities, relationship to respondent and multi-unit status did better in discriminating capture probabilities.

Also, the comparison between using SUDAAN and SAS is of interest. The parameter estimates were the same for practical purposes. However, the Wald statistics and the p-values differed. In the SAS models the variables *Male* and *Asian* were not statistically significant. The design effect used for SAS was constant for all variables, though the SUDAAN output clearly shows that the design effects vary for the variables.

VII. References

- Childers, Danny R. (1998): *The Design of the Census 2000 Dress Rehearsal Integrated Coverage Measurement*. DSSD Census 2000 Dress Rehearsal Memorandum Series, Chapter F-DT-2
- Hosmer, David W., Lemeshow, Stanley (1989): *Applied Logistic Regression*. John Wiley & Sons, New York.
- Waite, Preston Jay, and Hogan, Howard (1998): *Statistical Methodologies for Census 2000 Decisions, Issues and Preliminary Results*. Presented at the Joint Statistical Meetings, Session on Social Statistics, August 13, 1998, in Dallas, Texas.
- Wolter, Kirk M. (1986): *Some Coverage Error Models for Census Data*. Journal of the American Statistical Association, June 1986, Vol. 81, No. 394.
- SAS (1994): *SAT/STAT User's Guide, Version 6, Fourth Edition Volume 2*. The SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513
- Shah, B.V., Barnwell, B.G., and Bieler, G.S. (1997): *SUDAAN, User's Manual Volume II, Release 7.5*. Research Triangle Institute, Research Triangle Park, NC 27709

Privacy and Confidentiality Attitudes from the Census 2000 Dress Rehearsal
A paper presented at the American Statistical Association Annual Conference, August 8, 1999

Sara K. Buckley

Introduction

The cover of a recent edition of the Economist magazine proclaimed "The End of Privacy". In a series of articles, they highlighted the concerns and threats to privacy that the computer and internet revolution has help spur. These concerns about privacy and confidentiality may have a direct impact on the willingness of respondents to reply to surveys, and may have an impact on the propensity of respondents to return Census forms.

After the decline in response rates after the 1990 census, several people investigated the effects of privacy and confidentiality concerns on census participation. (Fay, Bates, and Moore 1991, Singer 1993). In preparation for Census 2000, the Census Bureau conducted a 1998 Dress Rehearsal Census in three sites, to test all of the operations planned for Census 2000. The dress rehearsal was conducted in Sacramento, CA, Columbia, SC and its surrounding counties, and the Menomonee, WI Indian Reservation.

As part of the Census 2000 Dress Rehearsal, the Census Bureau contracted with Westat, a survey research firm, to conduct a survey which would enable the Census Bureau to evaluate various aspects of public reaction to the Census. In the survey, there was a battery of questions that sought to investigate knowledge and attitudes about the Census, including those regarding privacy and confidentiality. These questions were similar to those that were asked in previous surveys conducted after the 1980 and 1990 Censuses, such as the Knowledge, Attitudes, and Practices Survey (KAP) and the Outreach Evaluation Survey (OES).

In this paper, privacy and confidentiality concerns are investigated as an explanatory factor in the actual mailback behavior in the 1998 Dress Rehearsal Census. Demographic predictors of privacy and confidentiality concerns are investigated and controlled for in the analysis.

Methodology

To address the evaluation research questions, Westat conducted a Random Digit Dial (RDD) telephone survey. The survey began approximately one week after the second (replacement) census form was mailed out. The surveys were conducted in the Sacramento and South Carolina test sites. The Menomonee site was not included as part of this survey because of its small population.

The Dress Rehearsal survey utilized a RDD sampling procedure with a goal of conducting 1,500 interviews in each of the two sites. Interviews were conducted between

May 1 and June 1 in South Carolina and between April 24 and June 7 in Sacramento. Census day was April 18, 1998 for the dress rehearsal. Interviews were conducted with the household member who usually handles the mail for the household. Both samples were post-stratified and weighted up to the test site population. In Sacramento, the response rate to the post-wave survey was 54%; in South Carolina it was 64%. This yielded 1,504 cases from Sacramento and 1,506 from South Carolina.

The Dress Rehearsal survey questionnaire consisted of several sections: media use, degree of civic participation, awareness of government agencies and programs, free recall of exposure to information about the census, aided recall about source of information, knowledge and attitudes about the census, aided recall of specific advertising, and census form receipt, handling, mailback behavior, and demographic information.

In this analysis, privacy and confidentiality are treated as two separate concerns. Confidentiality addresses respondents' concerns that data given to one agency is kept separate from another agency. Whereas privacy refers to wanting to keep information about oneself out of anyone's hands altogether. (Singer 1993)

Our outcome variable of interest—Census form mailback behavior—was measured two ways: from a self-reported survey question and from Dress Rehearsal Census Bureau records. The latter measure was made available by matching addresses obtained during the post-wave survey interview to the Census Bureau master list of addresses (the Master Address File). Because mail return questions are subject to social desirability effects, we wanted to use actual return records in order to have a more accurate model of behavior than self-reports.

Data/Design Limitations

As the survey was not designed specifically for privacy research purposes, this questionnaire did not have the number of privacy and confidentiality concern questions that previous research included. Therefore, the indexes developed were not as thorough as possible and were not able to examine all privacy and confidentiality concerns to the extent desired.

Because the survey was RDD, the findings reported from these data can only be generalized to telephone households in each site. It is estimated that 96.3% of households in the Sacramento test site have telephones and 90.5% have telephones in the South Carolina site (Evaluation E1a; 1998). This limitation is decreased somewhat by weighting the survey data to reflect the population, in each of the two sites, on income, race, and Hispanic origin.

The response rates were somewhat below industry standards for RDD surveys, especially in Sacramento. As a result of these response rates, the amount of non-sampling error in our data is potentially large. This leaves us to wonder if our respondents differed significantly from nonrespondents in terms of mailback behavior, attitudes towards the Census, and privacy and confidentiality concerns.

A final limitation to our data was the unanticipated loss of cases due to sampling outside the Dress Rehearsal geographic boundaries. This is critical because households that were interviewed for the Dress Rehearsal survey but located outside of the test sites were not mailed a census form. These out-of-scope cases were identified during the process of address geocoding and were especially prevalent in the Sacramento site.

Results

As noted earlier, the survey asked a battery of questions about census knowledge and attitudes similar to those in 1980 and 1990. For the sake of parsimony, exploratory factor analysis was done to determine the variables that would represent privacy and confidentiality concerns. Orthogonal factor analysis yielded four factors—of which, one dealt with privacy and another dealt with confidentiality. Appendix A shows the components of these indexes. This finding concurred with prior research that suggested that privacy and confidentiality were different concerns. (Singer, 1993).

Based on the factor analysis, separate indexes were made for privacy and confidentiality concerns—one for each dress rehearsal site. The items making the privacy index had an average factor loading score of 0.73 in South Carolina (range 0.71 to 0.75) and 0.63 in Sacramento (range 0.61 to 0.65). The items making the confidentiality index had an average factor loading score of 0.43 in South Carolina (range 0.43 to 0.44) and 0.52 in Sacramento (range 0.45 to 0.57). Cronbach's Alpha was computed for each index as a measure of index quality. For the privacy index, alpha was 0.76 in Sacramento and 0.78 in South Carolina. Similarly, the alpha for the confidentiality index was 0.70 in Sacramento and 0.71 in South Carolina. These alphas would indicate reliable indexes.

Table 1 shows the distributions for the indices. Note that a low score indicates little or no concern, while a high score indicates a moderate or high concern.

Table 1: Distributions of Privacy and Confidentiality Indexes in Sacramento

Index

Privacy

Confidentiality

Scores

Index (%)

Index (%)

0 (low)

57.6

57.0

1

17.3

23.7

2	
12.0	
19.2	

3	
13.2	
--	

Total	
100.0	
100.0	

Table 2: Distributions of Privacy and Confidentiality Indexes in South Carolina

Index	
Privacy	
Confidentiality	

Scores	
Index (%)	
Index (%)	

0 (low)	
42.4	
67.3	

1	
20.3	
20.0	

2	
14.9	
12.7	

3	
22.4	
--	

Total	
100.0	
100.0	

At this point, we investigated the possibility that there was a bivariate relationship between mailback behavior and privacy and confidentiality concerns. Table 3 shows

crosstabs between the Indexes described above, and the actual Census mailback behavior as matched to census records. This illustrates that there is not a relationship between privacy concerns and mailback behavior in the Sacramento area ($X^2=1.6$, d.f.=3, $p=0.663$). However, there is evidence of a relationship between confidentiality concerns and mailback behavior in Sacramento ($X^2=5.5$, d.f.=2, $p=0.064$). Yet, we see the opposite in South Carolina, where there is evidence of a relationship between privacy concerns and mailback behavior, and no significant relationship between confidentiality and mailback behavior.

Table 3: Relationship between mailback behavior and privacy and confidentiality concerns in Sacramento.

Index

Privacy

Confidentiality

Scores

Index (% return)

Index (% return)

0 (low)

64.96

67.53

1

65.70

64.32

2

67.54

56.47

3

59.52

--

X^2

1.585

5.503

d.f.

3

2

P

0.663

0.064

Table 4: Relationship between mailback behavior and privacy and confidentiality concerns in South Carolina

Index

Privacy

Confidentiality

Scores

Index (% return)

Index (% return)

0 (low)

70.73

70.09

1

70.69

71.76

2

63.15

60.60

3

62.56

--

X2

6.597

3.942

d.f.

3

2

p

0.086

0.139

At this point, we performed linear regression models to investigate demographic characteristics that contribute to privacy and confidentiality concerns. Here, we used prior research (Singer, 1993) as a guide to which factors might be relevant. Based on previous research, we investigated the effects of race/ ethnicity, education, gender, presence of children at home, and whether or not respondents were born in the US. The

rationale was that foreign-born respondents might be more concerned about immigration laws, and thus have more concerns. Language spoken at home was not included in the models because of its strong correlation with being foreign-born. The presence of both variables in the model introduced more problems than the value it added.

Table 5: Linear Regression Coefficients of Demographic Characteristics for Privacy and Confidentiality Indexes in Sacramento

Privacy

Standard

Variable
Coefficient
Error

Age
0.044 *
0.023

Education
-0.161 ***
0.033

Gender (Female)
-0.017
0.075

Foreign Born
0.124
0.121

Race

Black
0.366 ***

0.105

Hispanic

0.202 *

0.111

Asian

0.526 ***

0.126

Other

-0.060

0.173

Children at home

0.091

0.012 -

Household Income

-0.054 ***

0.081

Adj. R2

= 0.1496

*p<0.10, **p<0.05, ***p<0.01

N=772

Confidentiality

Standard

Variable

Coefficient

Error

Age

0.004

0.017

Education

0.030

0.026

Gender (Female)

-0.011

0.058

Foreign Born

-0.094

0.092

Race

Black

0.185 **

0.082 -

Hispanic

-0.049

0.086

Asian

-0.037

0.095

Other

0.469 ***

0.131

Children at home

0.013

0.062

Household Income

0.007

0.009

Adj. R2

= 0.0232

*p<0.10, **p<0.05, ***p<0.01

N=772

In Sacramento, it appears that those respondents with higher levels of education or

household income, all else being equal, tend to have fewer privacy concerns. Conversely it appears that older respondents have more privacy concerns. Blacks, Hispanics, and Asians/Pacific Islanders tend to have higher levels of privacy concerns relative to whites, the omitted category. Looking at the confidentiality issues, it appears that only Blacks and Other Races have more confidentiality concerns than whites.

Next, the same demographic characteristics are investigated for the South Carolina site.

Table 6: Linear Regression Coefficients of Demographic Characteristics for Privacy and Confidentiality Indexes in South Carolina

Privacy

Standard

**Variable
Coefficient
Error**

Age
0.071 ***
0.027

Education
-0.101 ***
0.030

Gender (Female)
-0.034
0.077

Foreign Born
0.307
0.246

Race

Nonwhite
0.390 ***
0.077

Children at home

-0.060

0.077

Living in City

-0.210 ***

0.012

Household Income

-0.052 ***

0.078

Adj. R2

= 0.1256

*p<0.10, **p<0.05, ***p<0.01

N=910

Confidentiality

Standard

Variable

Coefficient

Error

Age

0.040 **

0.017

Education

0.021

0.019

Gender (Female)

-0.036

0.049

Foreign Born

0.170

0.157

Race

Nonwhite

0.024

0.049

Children at home

0.013

0.049

Living in City

0.098 **

0.048 -

Household Income

0.018**

0.008

Adj. R2

= 0.0190

*p<0.10, **p<0.05, ***p<0.01

N=910

As in Sacramento, the South Carolina sample seems to have some of the same concerns with privacy. Once again, it appears that those respondents with higher levels of education or household income, all else being equal, tend to face fewer privacy concerns. Older respondents face more privacy concerns. Races other than White, also tend to have higher levels of privacy concerns relative to whites, the omitted category. Because the South Carolina site comprised urban and rural areas, we were able to separate out respondents living within the city of Columbia, to investigate if that had any effect on privacy and confidentiality concerns. Here, we see that those respondents living in the city have fewer privacy concerns.

However, what is interesting about the South Carolina regressions, is that when looking at the confidentiality predictors, we see that those living in the city have a higher confidentiality concern. It would appear that city dwellers do not mind if people have information about them, but are concerned about what is done with it. Also, we note that older respondents have more confidentiality concerns, as do respondents with higher incomes.

Mailback Behavior

Although the bivariate relationship between mailback behavior, privacy, and confidentiality issues was looked at earlier, it is important to look at the effect privacy and confidentiality attitudes have on mailback behavior, while controlling for other demographic variables, such as those in the previous section. In this context, we look at logistic regression models, using both the PROC LOGISTIC and PROC GENMOD functions in SAS. Although only the best models are reported, many models, both with and without interaction terms, were investigated.

Sacramento

Table 7: Census Form Mailback Behavior and Demographic Variables for Sacramento Test Site

Parameter
Standard

Estimate
Error

Civic Participation *

0.209

0.108

Privacy

-0.155

0.109

Confidentiality **

-0.236

0.118

Race:

Black *

-0.453

0.262

API
-0.437
0.347

Hispanic *
-0.569
0.324

Other
-0.129
0.393

Expecting Form **
0.466
0.181

Age ***
0.673
0.125

Hispanic*Privacy *
0.383
0.220

API*Privacy **
0.555
0.240

*p<0.10, **p<0.05, ***p<0.01
N=780

As seen in table 7, there are several factors that impact the odds of returning a census form in the mail. Civic participation, expecting a census form in the mail, and age all contribute positively to the probability that a respondent will return the census form. However, as previous research showed, race has an impact on the odds of returning a form. All else being equal, Blacks are 37% less likely to return their forms relative to whites, and Hispanics are 44% less likely to return their forms relative to whites. While privacy concerns, on their own, do not have a significant effect, we do see that confidentiality concerns do. For each increase in confidentiality concern on the index, respondents are 21% less likely to return their forms.

There is an interesting interaction in Sacramento with Hispanics and privacy and API groups and privacy. It appears that Hispanics with privacy concerns are 25% more likely to mailback their forms, and API's with privacy concerns are 49% more likely to return their forms. While this seems like a strange result initially, perhaps these groups respond because of the nature of the data collection agency.

South Carolina

Now, we look at the same analysis for the South Carolina test site. Table 8 shows the output. Once again, only the best model is reported here.

Table 8: Census Form Mailback Behavior and Demographic Variables for the South Carolina test site.

Parameter
Standard

Estimate
Error

Civic Participation ***
0.305
0.108

Privacy
0.013
0.077

Confidentiality *
-0.217
0.125

Race:

Nonwhite ***
-0.613
0.181

Expecting Form ***
0.577
0.176

Age ***
0.345
0.127

Education **
0.327

0.150

Income **

-0.248

0.125

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

N=851

In South Carolina civic participation, age, education and expecting a form are all positively related to returning a census form. Here, income has a negative relationship with returning a form, which is a rather counter intuitive result. It is possible that once other factors, such as race, age etc., are controlled for, individuals with higher income levels are less likely to return their Census forms. While privacy, both on it's own and interacting with other variables, is not significant, confidentiality is significant. Individuals with confidentiality concerns are roughly 20% (odds ratio=0.805) less likely than those without privacy concerns to mail back their Census forms. As well, race plays an important factor in mailing back the form. Here, Races other than White were 44% (odds ratio=0.542) less likely to mail back their census form than whites, the control group.

Conclusions

This research was done to determine what, if any, impact privacy and confidentiality concerns had on the mailback rates for the 1998 Census Dress Rehearsal. It is important to note that there were limitations to the analysis due to the limited measures available for the privacy and confidentiality measures. These findings may have been much different had there been more information available for the indexes.

Although the privacy index does not have a significant effect on mail back rates, the confidentiality index does. Respondents with stronger confidentiality concerns were less likely to mail back their form. This is true for both dress rehearsal sites.

The demographic predictors of privacy and confidentiality were not very good. Although we were able to account for roughly 15% of the variance in privacy concerns in Sacramento, and 13% in South Carolina, privacy did not seem to be a significant predictor in mailback behavior. Further, we were only able to account for 3% of the variance concerns for confidentiality in Sacramento, and only 2% in South Carolina. Clearly, demographic variables were not good predictors of confidentiality concerns. Although some individual variables, particularly the race and age of the respondent, are good predictors of the indexes, the lack of explained variance suggests that there are other important predictors besides demographic characteristics that are critical in explaining privacy and confidentiality concerns.

If the predictions of the Economist are true, and confidentiality and privacy concerns become more salient to the American public, the effect on mailback rates could be

significant. Responding to the public's privacy and confidentiality concerns could be an integral strategy in order to keep census mail return rates from dropping. Clearly, more research needs to be conducted about the links between privacy and confidentiality concerns and census mail return.

References

Bates, N., Buckley, S.K. (1998)

"Effectiveness of the Paid Advertising Campaign: Reported Exposure to Advertising and Likelihood of Returning a Census Form," PRED Census 2000 Dress Rehearsal Evaluation Memorandum Series #E1b.

Evaluation E1a by Roper Starch Worldwide (1998)

"Promotion Evaluation: Evaluating the Effectiveness of Paid Advertising in the Census 2000 Dress Rehearsal in Sacramento and South Carolina." Report prepared for Bureau of the Census, U.S. Department of Commerce. December 3, 1998.

Fay, R.E., Bates, N., and Moore, J. (1991)

"A Lower Mail Response in the 1990 Census: A Preliminary Interpretation," US Census Bureau 1991 Annual Research Conference proceedings, pp. 3-32.

Kerwin, J., Edwards, S. (1996)

"The 1996 Study of Public Attitudes Toward Administrative Records Use: Final Report," Report prepared for US Census Bureau, US Department of Commerce, December 18, 1996.

Moore, J. (1982)

"Evaluating the Public Information Campaign for the 1980 Census: Results of the KAP Survey," US Census Bureau Results Memorandum No. 31.

Singer, E., Mathiowetz, N.A., Couper, M.P. (1993)

The Impact of Privacy and Confidentiality Concerns on Survey Participation: The Case of the 1990 US Census, Public Opinion Quarterly Vol. 57, p. 465-482.

Appendix A: Survey Questions—Construction of Indexes

Privacy Index

Is the Census Used To:

Q19_A "Check up on whether people are paying their income taxes"

Q19_C "Locate people living in the country illegally"

Q19_E "Police and FBI use the census to keep track of people who break the law"

Confidentiality Index

Q17_A "The Census Bureau's promise of confidentiality can be trusted"

Q17_D "The Census Bureau would never let another government agency see my answers to the Census"

Appendix B: Definition of Independent Variables

Age: Measured as a categorical variable 18-34, 35-54, 55+ for logistic regression. More categories were used in multiple regression: 18-24, 25-34, 35-44, 45-54, 55-64, 65+.

Education: Categorical variable with following categories: Less than high school, High School and some college, College degree and higher.

Gender: Dichotomous variable: assigned 1 if female, 0 if male.

Foreign Born: Assigned 1 if answered yes to Q40: Were you born in the United States, 0 otherwise.

Children at Home: Assigned 1 if answered yes to Q36: Do you have children in school who are under 18 living at home with you?

Household Income: For Logistic regression categorical variable: Less than 25,000, between 25,000 and 50,000 and greater than 50,000. For multiple regression analysis, continuous variable ranging from under 15,000 to greater than 100,000.

Race: Respondents who answered yes to Q38: Are you of Hispanic origin or descent?" were coded as Hispanic, regardless of their racial choice. Respondents were also asked Q39: Which of these categories best describes your race?" Respondents in South Carolina were coded as White, or Races other than white, based on the distribution of the racial breakdowns. In Sacramento, we used the following classifications: White non-Hispanic, Black non-Hispanic, Hispanic, Asian/Other Pacific Islander non-Hispanic and Other non-Hispanic.

Expecting a Form: Respondents were asked whether they expected a census form in the mail. Those answering yes were coded as one, those answering no or not asked were coded as 0.

Civic Participation: Civic participation was defined by a battery of yes/no questions (Q3) inquiring about work for a political party or candidate, non political volunteer work, voting in the last presidential and local elections, membership in a PTA, religious organization, civic club or community organization, and membership in a union. Answers to these questions were summed with the highest score equal to 8 and lowest score equal to 0. This score was then categorized in to four levels (0 activities, 1 4 and 5 or more).

Response rates were calculated using 1998 guidelines of the American Association for Public Opinion Research.

See Appendix B for description of variables.

ASSESSING THE QUALITY OF THE INITIAL MASTER ADDRESS FILE FOR CENSUS 2000

Joseph Burcham, Diane Barrett, U.S. Census Bureau, Planning, Research, and Evaluation Division
Joseph Burcham, U.S. Census Bureau, Room BH118-2, Washington, D.C. 20233

Key Words: QIP, MAF, coverage errors, geocoding errors

Introduction

The Master Address File, or MAF, is a file of residential addresses that will be used for Census 2000 and also will be maintained as a sampling frame after Census 2000.

At this point in time, the Bureau has used four different sources of addresses to update the MAF in areas where the Census Bureau will use mailout/mailback enumeration in Census 2000. The four sources were:

- 1990 Address Control File (ACF)
- November 1997 Delivery Sequence File (DSF) from the U.S. Postal Service
- September 1998 DSF
- Address files from the Block Canvassing operation

The ACF is the file of addresses collected during the 1990 Census. The DSF is maintained by the U.S. Postal Service and contains more up-to-date information about residential and non-residential addresses that receive mail. In the Block Canvassing operation, field representatives traveled to the mailout/mailback areas to provide additional updates to the MAF.

The Initial MAF, which is the version of the MAF we measured in this study, consisted of the ACF and the November 1997 DSF.

Because the Census Bureau must have the ability to geographically locate each address, each address on the MAF is assigned, or geocoded, to a census block.

Goal of the Evaluation

Census Bureau staff designed the 1998 Quality Improvement Program (QIP) to measure the effectiveness of the Initial MAF in accurately reflecting existing housing units as of April 1, 1998.

The MAF will eventually become the single source of addresses that the Bureau will use to conduct Census 2000 as well as other surveys. The point of measuring an early version of the file is that we wanted to see how good the two sources of addresses used to create the file were. By understanding the impact of these two sources, we could get an indication of the amount of coverage improvement that was needed on the file.

We accomplished the goal by producing national level and census division level ratio estimates of coverage errors and coding errors on the MAF.

We had data suggesting that the quality of the Delivery Sequence File varied among postal service areas. During early stages in planning, we explored the possibility of producing estimates for each postal service area. Due to the fact that postal service areas cross county boundaries, we were unable to assign our sample counties to a specific postal service area. Therefore, we produced estimates for the level of census geography that is closest to postal service area, which is census division.

The five errors that we were specifically interested in measuring were:

- *undercoverage error* (an existing residential address is missing from the MAF)
- *overcoverage error* (a non-existing residential address is included on the MAF)
- *geocoding error* (an existing residential address is coded to the wrong census block on the MAF)
- *ungeocodable error* (an existing residential address is on the MAF, but not coded to a

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau Publications. This report is released to inform interested parties of research and to encourage discussion.

census block at all)

- *non-residential coding error* (an existing residential address is incorrectly coded "non-residential" on the MAF)

The 1998 QIP Operation

The QIP team modeled QIP methodology after the Census Bureau's 1996 Integrated Coverage Measurement (ICM) methodology. ICM is currently known as the Accuracy and Coverage Evaluation, and was designed to measure the coverage of people and housing units in the Census. To test the operational feasibility of using the Integrated Coverage Measurement methodology for QIP, the team conducted a pilot study in 1997 in six counties. With a few modifications, the methodology proved to be effective in measuring the coverage of housing units on the MAF.

The operation consisted of the following steps:

- selecting a stratified, two stage cluster sample of areas to be used in the study (the first stage being a sample of counties and the second stage being a sample of blocks)
- creating the Independent Listing (IL), which was a current list of all housing units existing in the blocks.
- Matching the IL to the MAF (to evaluate the MAF)
- computing estimates of MAF errors using the match codes
- computing standard errors using stratified jackknife replication

Sample Design

We designed our sample to give us coefficients of variation of about 10 to 15%.

The universe consisted of all counties that contained some areas classified by mailout/mailback enumeration.

Mailout/mailback enumeration areas are areas consisting of households that will receive their Census 2000 forms in the mail. These areas consist of primarily city-style addresses.

To stratify the universe of counties, we first grouped the

counties by Census Division. Within each Division, we assigned counties in the universe to one of four "growth" cells. The four "growth" cells were called:

- Low/Low
- Low/High
- High/Low
- High/High

We created these cells by comparing housing unit growth estimates on the Delivery Sequence File to the Census Bureau's housing unit growth estimates. Sometimes the estimates agreed and sometimes they did not. We set up the 4 cells to reflect the level of agreement. For example, the low/low cell identifies low housing unit growth on the Delivery Sequence File and low housing unit growth according to the Census Bureau, etc.

Nine divisions times 4 "growth" cells resulted in 36 "growth" strata nationwide.

Within each stratum, we selected 3 counties systematically proportional to the size of the county. Selecting 3 counties in each stratum resulted in 108 total counties.

We required a total of about 170,000 residential addresses in sample. We allocated that sample size to the counties in order to achieve a self-weighting design within each "growth" stratum.

Within each sample county, we selected a sample of blocks. The blocks were grouped into 4 "size" strata. A block is assigned to one of these strata based on the estimated number of units in the block. The four size strata were:

- 0-2 HUs
- 3-19 HUs
- 20-79 HUs
- 80+ HUs

We allocated the county sample size proportional to the count of HUs in the bottom three size strata. Then, in each of the three strata, we selected the required number of blocks with equal probability.

For the 0-2 strata, we wanted to select as few of these blocks as possible but also minimize the potential impact on variance of high housing unit growth in these blocks. So, we selected twelve blocks with equal probability in each of these strata.

The Independent Listing

To create the Independent Listing (IL), field representatives traveled to the blocks and listed all residential units that existed on April 1, 1998. It is assumed that this listing is more current than the MAF. So, we evaluated the MAF by comparing, or matching, the IL to the MAF.

Matching

In matching, whenever a residential IL address matched to a residential MAF address in the same block, this confirmed an existing housing unit.

Whenever an address on one list did not match an address on the other list, or whenever two addresses matched but the MAF address was coded in error, we potentially identified one of the errors we were measuring.

The first type of matching we did was computer matching. All IL addresses were residential and geocoded to one of the sample blocks. But on the MAF, we matched to addresses regardless of whether they were residential or non-residential or geocoded or ungeocoded. Also, on the MAF we matched to addresses coded to the sample block but also to addresses in the zip codes surrounding the sample blocks. The point of matching to a larger number of addresses on the MAF is that we have a better chance of identifying MAF coding errors.

Because the computer match was not perfect, we had several followup operations for the purpose of finding more matches and verifying the existence of units.

These followup operations were:

- Before Followup Review - which was clerical matching that occurred right after the computer match
- Field Followup - where field representatives traveled back to the sample blocks for the cases that were still unresolved
- After Followup Review - where the clerical matchers assigned final match codes

In field followup, we required field representatives to determine whether or not housing units existed for the addresses in question. There were some situations where we had an address that corresponded to an existing unit, but the address was incorrect. In these situations, some field representatives answered "the address does not exist," which was later interpreted as "the unit does not

exist" in After Followup Review.

In future studies of this nature, we should prevent these problems by making clearer procedures and/or different wording on the followup forms.

Estimation

We used the final match codes to produce ratio estimates for each of the five MAF errors that we were interested in.

When final match codes were being assigned, some addresses were still unresolved. An address could be unresolved because:

- you don't know whether it refers to an existing residential unit, or
- you don't know the correct block

To determine the impact of unresolved cases on the estimates, we computed the estimates two different ways, by:

- Excluding the unresolved cases, and
- Including the unresolved cases and assuming a worst-case scenario

For the most part, each worst-case scenario estimate was worse than its corresponding no-unresolved estimate by only about half of a percentage point. We decided to be conservative and present the worst-case scenario estimates.

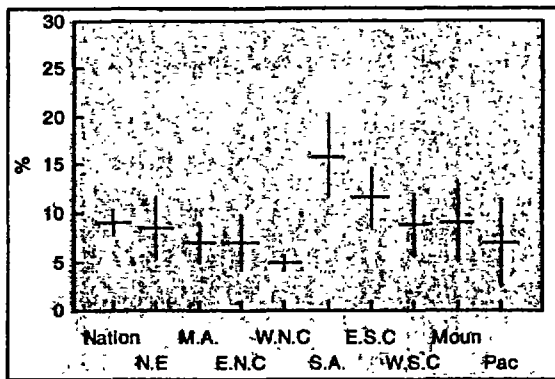
The Undercoverage Estimate

Undercoverage simply means a unit was missing from the MAF but we found it on the ground.

Specifically, this estimate is defined as: the percentage of existing housing units in the sample blocks that are missing from the MAF.

Figure 1 shows 90% confidence intervals for the national and census division level undercoverage estimates.

Figure 1. Undercoverage



The national undercoverage estimate (on the far left) is 9.1%. The confidence interval ranges from 7.8% to 10.3%.

The census divisions are abbreviated on the graph. The abbreviations of the divisions, with their names, are:

N.E. - New England
M.A. - Middle Atlantic
E.N.C. - East North Central
W.N.C. - West North Central
S.A. - South Atlantic
E.S.C. - East South Central
W.S.C. - West South Central
Moun - Mountain
Pac - Pacific

The undercoverage estimate ranges from about 5% in the West North Central division to about 16% in the South Atlantic division. The undercoverage rate in the South Atlantic division is significantly higher than the undercoverage rates in four other divisions.

In general, southern areas of the United States tend to display higher undercoverage than northern areas.

As stated before, all estimates that we present are worst-case scenario estimates. Most of the worst-case scenario estimates are worse than their corresponding no-unresolved estimates by only about half of a percentage point. The South Atlantic division is an exception. The undercoverage estimate in this division ranges from a no-unresolved estimate of 14.8% to a worst-case estimate of 15.9%.

The Overcoverage Estimate

Overcoverage means a unit was on the MAF but we did not find it on the ground.

Specifically:

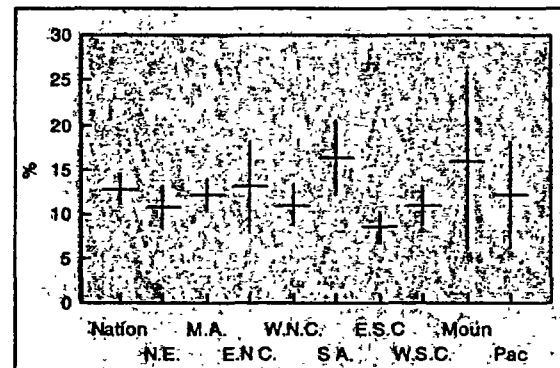
This estimate is the percentage of MAF addresses coded to the sample blocks that should not be coded to the sample blocks.

An overcoverage addresses could be:

- an address that refers to a housing unit that exists outside of the sample blocks
- an address that does not refer to an existing housing unit at all, or
- a duplicate of another residential MAF address

Figure 2 shows 90% confidence intervals for overcoverage.

Figure 2. Overcoverage



The national overcoverage estimate is 12.8%, with a confidence interval ranging from 11.1% to 14.6%.

Overcoverage ranges from about 8.5% in the East South Central division to about 16% in the South Atlantic division. These two rates are the only rates that are significantly different.

The Geocoding Error Estimate

Geocoding error means we found the unit on the ground but it was coded to the wrong block on the MAF.

Specifically, this estimate is the percentage of MAF housing units existing in the sample blocks that are geocoded in error.

When conducting a study of geocoding error based on sample blocks, there are different types of geocoding errors to consider:

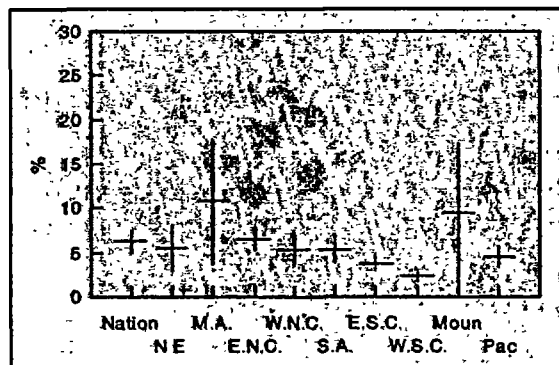
- *geocoding error of exclusion* - a housing unit exists inside a sample block but is incorrectly excluded from the sample block on the MAF (it is coded to the wrong block on the MAF)
- *geocoding error of inclusion* - a housing unit is incorrectly included in a sample block on the MAF, but exists outside of the sample block (it is coded to the wrong block on the MAF)

Another thing to consider is the area in which one searches for geocoding errors.

Because of limited resources, the only type of geocoding error we measured was geocoding error of exclusion. During the computer match, we searched for these errors within the sample blocks but also within zip code on the MAF. These types of geocoding errors are the only types reflected in our geocoding error estimate.

Figure 3 shows 90% confidence intervals for geocoding error.

Figure 3. Geocoding Error



Our national estimate of geocoding error is 6.2%, with a confidence interval ranging from 4.9% to 7.5%

Geocoding error ranges from about 2.5% in the West South Central division to about 11% in the Middle Atlantic division. The West South Central estimate is significantly lower than the estimate in five other divisions.

The geocoding error estimate in the South Atlantic division is also an exception to the 0.5% rule. Geocoding error in this division ranges from a no-unresolved estimate of 4.7% to a worst-case estimate of 5.4%.

The Ungecodable Match Rate Estimate

This estimate is a measurement of the extent that we

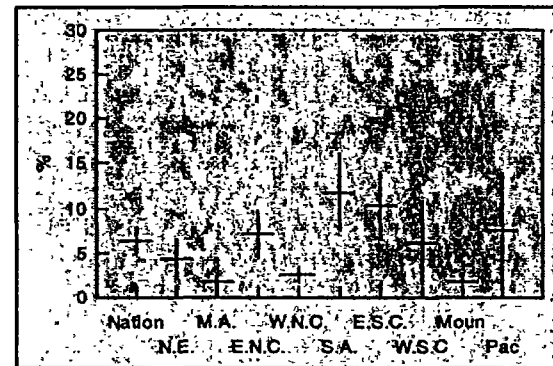
found units on the ground that were ungecodable on the MAF.

Specifically:

This estimate is the percentage of MAF housing units existing in the sample blocks that are ungecoded.

Figure 4 shows 90% confidence intervals for ungecodable match rate.

Figure 4. Ungecodable Match Rate



The national estimate is 6.4%, with a confidence interval ranging from 4.9% to 7.9%.

Ungecodable Match Rate ranges from about 2% in the Middle Atlantic division to about 12% in the South Atlantic division. The South Atlantic estimate is significantly higher than the estimate in four other divisions.

The Non-residential Coding Error Estimate

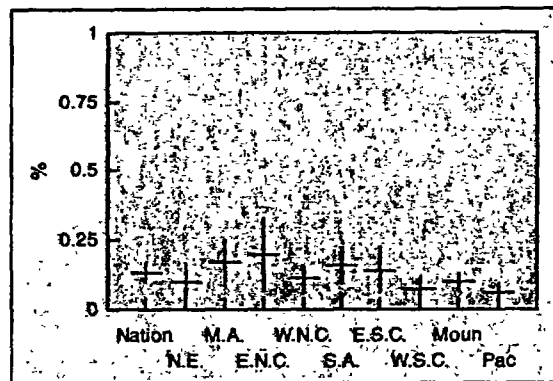
Non-residential coding error means we found a unit on the ground but it was incorrectly coded non-residential on the MAF.

Specifically:

This estimate is the percentage of residential MAF units in the sample blocks that are incorrectly coded non-residential.

Figure 5 shows 90% confidence intervals for non-residential coding error. Notice that all of the graphs shown previously had a y-axis ranging from 0 to 30%. This graph has a y-axis ranging from 0 to 1%.

Figure 5. Non-residential Coding Error



The national estimate is only 0.13%, with a confidence interval ranging from 0.1% to 0.16%.

Non-residential coding error is less than a fourth of a percent in every census division.

Relationship Between the Estimates

One relationship between the estimates that is worth mentioning is that of coding errors vs. coverage errors. We were successful in distinguishing between coverage errors (undercoverage/overcoverage) and coding errors, to the extent that we located coding errors within zip code on the MAF. Without this distinction, cases that were actually on the MAF but coded in error would appear to us as missing from the MAF, or in other words, undercoverage. So, our undercoverage estimate is lower and more accurate than it would have been without accounting for these coding errors.

Conclusions

In theory, the MAF should contain all residential units in the nation. Because most of our estimates are fairly high, this confirms the need for significant coverage improvement on the MAF prior to Census 2000.

The South Atlantic division appears to contain more MAF deficiencies than any other division. It shows the highest undercoverage rate, overcoverage rate, and ungeocodable match rate. These deficiencies may be due to the quality of the Delivery Sequence File in the Southeast and Mid-Atlantic postal service areas.

The lowest undercoverage rate in any division is 7% and lowest overcoverage rate is 8.5%. Because even the lowest estimates are relatively high, every census division shows a need for coverage improvement.

Because all of our estimates of non-residential coding error were so low, we do not believe this error is a major concern for MAF building.

As we approach Census 2000, one of the most important address building operations is the Block Canvassing operation. Again, the purpose of Block Canvassing is to improve the quality of the MAF in all mailout/mailback areas of the nation. Because of our relatively high estimates of errors on the Initial MAF, we believe the Block Canvassing operation is critical in ensuring the highest possible quality of the MAF.

In past censuses, field work has always been essential in creating the address file. Because the Initial MAF for Census 2000 was created without any field operations, it may not be a big surprise that we measured such a high number of errors. In addition to the Block Canvassing operation, several other field operations are being planned to update the MAF prior to Census 2000.

References

Barrett, Diane F. (1999), "The 1998 Master Address File Quality Improvement Program," internal memorandum for Robert Marx, Bureau of the Census.

Barrett, Diane F. (1999), "The 1997 Master Address File Quality Improvement Program Pilot Study," internal memorandum for Robert Marx, Bureau of the Census.

Bureau of the Census (1997), "Draft Sections of the Census 2000 History," internal memorandum, Bureau of the Census.

Bureau of the Census (1999), "Program Master Plan: Census 2000 Block Canvassing Operation," internal memorandum, Bureau of the Census.

(Need DSF source telling about November 97 and September 98)

(Need ASA paper for ICM)

CENSUS 2000 DRESS REHEARSAL METHODOLOGY AND INITIAL RESULTS

Rajendra P. Singh, Patrick J. Cantwell, and Donna L. Kostanich, Bureau of the Census*
Patrick J. Cantwell, Bureau of the Census, Washington, DC 20233-7600
patrick.j.cantwell@ccmail.census.gov

Key Words: Nonresponse follow-up, undercoverage, dual system estimation, post-stratification, undercount.

1 Introduction

The U.S. Bureau of the Census conducts a census of population and housing units every ten years. Counting every person is important as the census is used to apportion the seats in Congress among the states, to allocate billions of dollars of federal funds, and for future planning by federal and private agencies. Yet the census has been afflicted by two serious problems in recent decades. First the price of the census has increased markedly. The 1990 Census cost \$2.6 billion; in constant 1990 dollars, the four censuses from 1960 to 1990 cost \$9, \$11, \$20, and \$25 per household, respectively (National Research Council 1995). The 2000 Census is projected to cost much more. As will be discussed in the next section, much of this increase is due to the fact that fewer people have been returning their census questionnaires.

The second problem involves the census counts themselves. For the past six censuses, the Bureau has evaluated errors in census counts—measuring people who were missed and others who were enumerated erroneously. The evaluations have shown a net undercount for the total population, and a differential net undercount among demographic groups. For example, the Bureau estimated that in the 1990 Census 1.6 percent of all people were missed, but that 4.4 percent of all Blacks were missed, as well as 5.0 percent of all Hispanics.

To address these problems, during the 1990's the Bureau of the Census developed methods for conducting the census that rely on sampling and estimation procedures. First, instead of visiting all housing units that did not respond, our plan was to select a large sample of them and visit only those in the sample. A second component of the new procedures was called the Integrated Coverage Measurement (ICM). This quality check survey and operation, similar to the Post Enumeration Survey conducted after the 1990 Census but larger in scope, would allow us to estimate the net undercount for various demographic groups and to adjust the census counts—thereby eliminating or greatly reducing the differential net undercount among the groups.

In 1998 the Census Bureau conducted a dress rehearsal in three sites. According to an agreement

between the Congress and the Department of Commerce, we applied the planned sampling techniques in two of the sites—Sacramento, California, and Menominee County, Wisconsin. In the third site, the city of Columbia, South Carolina and eleven surrounding counties, sampling procedures were not used. However, a post-enumeration survey was conducted there to measure the net undercount. This paper discusses the methodology used in the Dress Rehearsal and presents a brief summary of selected results in the three dress rehearsal sites.

On January 25, 1999, the Supreme Court ruled against the use of sampling in the 2000 Census for the apportionment counts, but left open its use for all other purposes. Thus, the Census Bureau will follow up all nonresponding households in 2000, but will conduct a quality check survey to estimate and adjust for the undercount—for purposes other than apportionment.

In Section 2 we present the methodology and results for the initial phase of the census—including sampling for nonresponse follow-up. Section 3 provides a discussion and some results from the ICM phase.

2 The Initial Phase

Since the 1970 census, most people have been mailed or given a questionnaire, and asked to mail it back. To each household that does not return a questionnaire by mail, the Bureau sends an interviewer to collect the information. This 100% follow-up of the nonresponding households has been a major factor in the increase in census cost as more and more households require follow-up with a personal visit. The response rate for the mail questionnaire declined from 78% in the 1970 Census, to 75% in 1980, to 65% in 1990 (National Research Council 1995). To address the issue of declining response rates and increased costs, the Bureau developed a plan to sample the nonresponding households.

To begin the initial phase of the Dress Rehearsal the Census Bureau prepared a list of addresses in the three sites and delivered census questionnaires by mail or in person to all housing units on the list. The U.S. Postal Service returned some of the questionnaires as "undeliverable as addressed." In addition, we did not receive questionnaires back from many of the remaining housing units. The Bureau conducted a field follow-up of nonresponding housing units and postal returns. As is seen below, the procedures differed according to the site.

2.1 Sampling for Nonresponse Follow-up (NRFU)

In Sacramento, California, from those households that did not return a census form, we selected a sample and conducted a personal interview. Although the Bureau applied sampling techniques in Menominee County, Wisconsin, during the Dress Rehearsal, most of the Menominee site is an Indian reservation. On Indian reservations our sampling plans called for personal follow-up visits to *all* the nonresponding housing units. In South Carolina, we also conducted follow-up with all nonresponding units—as is the general rule under a traditional census.

In Sacramento, the sample of nonresponding housing units was selected separately in each census tract. The goal was to achieve a completion rate of 90% or higher in each tract. For example, if the initial response rate in a tract was 70%, then the sampling rate among nonrespondents would be 2 in 3. In general, if the initial response rate was 85% or higher, we sampled 1 in 3. Thus the sampling rate depended on the initial response rate, and varied by tract; as the response rate increased, our sampling rate in that tract generally decreased. Note, however, that we followed up all nonrespondents in blocks selected for the Integrated Coverage Measurement sample. (See Section 3.1.)

Immediately after the cut-off for mail returns, the nonresponding housing units were sorted within the tract by geography and form type (long vs. short form). Then a systematic sample was selected. This ensured that the sample was distributed evenly across the tract.

The population characteristics of the remaining nonresponding households—those not selected in the sample—were imputed using a hot-deck procedure based on information collected from sampled nonrespondents in the same census tract. (This was also true for housing units selected in the NRFU sample that were nonrespondents again in the follow-up.) The procedure was designed to reduce bias in estimation and to ensure that the hot-deck population estimates agree in expectation with simple weighted estimates at the tract and higher levels of geography. Occasionally a census form was returned after NRFU sample selection. The information from these forms was not discarded; rather an appropriate adjustment to the estimation methodology was made to accommodate late forms.

2.2 Sampling the Undeliverable-As-Addressed (UAA) Vacant Returns

In the Sacramento site, we selected a sample of the UAA vacant housing units for personal visits to check whether the housing unit was actually vacant. As was the case with nonresponse follow-up, there was no UAA

sampling operation in Menominee or South Carolina. In Menominee, Census Bureau staff initially left questionnaires at all existing housing units; because the U.S. Postal Service did not deliver questionnaires, there were no UAA vacant returns there. As was mentioned earlier, in South Carolina we conducted a traditional census.

The Bureau selected a systematic sample of UAA vacant units as identified by the U.S. Postal Service at a 3-in-10 rate. This rate was the same in each tract, regardless of the number of returns. Before selecting the sample, the vacant returns were sorted by geography and form type (long vs. short) within an eligible tract. The sampling for NRFU and UAA vacants was done simultaneously but independently within separate sampling strata for NRFU units and UAA vacants. The characteristics for UAA vacant units not selected into the sample were imputed based on sample cases in the same census tract.

2.3 Results of Sampling for NRFU and UAA in Sacramento

Table 1 below presents some results for the initial phase in the collection blocks in Sacramento (Memo. [1]).

Table 1. Sacramento, CA: Housing Unit Frequencies in Collection Blocks

Occupied Housing Units	Total	% of Total
Total	151,732	100.00%
Total respondents	138,271	91.1%
By mail	90,156	59.4%
In NRFU sample	48,115	31.7%
Imputed	13,461	8.9%

On a person basis, in Sacramento 24,930 people were imputed into nonresponding housing units based on the NRFU sampling and estimation operations. Similarly, 2,409 people were imputed into UAA addresses. As noted earlier, we followed up all nonresponding units in the Menominee and Columbia sites.

3 Integrated Coverage Measurement (ICM) and Post Enumeration Survey (PES)

To reduce the differential net undercount, the Bureau conducted the ICM in Sacramento and Menominee and

adjusted the initial-phase counts. Essentially the same design was implemented in South Carolina in what was called the Post Enumeration Survey (PES). The goal in South Carolina, however, was to measure the net and differential net undercount without making adjustments to the counts. *Unless otherwise specified*, whenever we refer to the ICM, our statements refer to the ICM conducted in Sacramento and Menominee *as well as* the PES in South Carolina.

3.1 Sampling in the ICM (PES)

The ICM sample was a stratified systematic sample of clusters of geographically contiguous housing units. To select the sample, we first formed block clusters by combining adjacent blocks with at least three housing units but no more than 79. All other blocks were defined as block clusters by themselves.

Next we formed sampling strata by grouping block clusters into homogeneous groups based on the 1990 Census demographic characteristics of the block clusters. The stratum definitions corresponded to major demographic groups such as the proportion of certain race or ethnic groups and the proportion of renters—known to be traditionally undercounted. These strata were formed within each group of clusters based on their sizes—small (1-2 housing units), medium (3-79 housing units), and large (80 or more housing units) clusters. In the third step we selected sample clusters using proportional allocation and systematic sampling.

In the fourth step Census Bureau field staff listed the sample block clusters independently, that is, without the use of any census address lists. If a selected cluster had fewer than 80 housing units in the independent listing, all were retained for ICM interviewing. If the cluster had 80 or more housing units it was divided into segments and one or more of these segments were selected randomly. This was done to make the interviewer work load more efficient and to improve the efficiency of the design by reducing the clustering effect. Other adjustments to the sample were also applied to bring field listing workloads in line with expectation. A sample of block clusters was selected for each site. The summary statistics of the ICM samples by site are given in Table 2.

3.2 Interviewing Results

Census Day in the Dress Rehearsal was April 18, 1998. ICM telephone interviewing started May 15, and the personal visit (door-to-door) interviewing started on June 2; interviewing was completed on September 3, 1998. Computer Assisted Personal Interviewing (CAPI) was used for the ICM sample households.

The "Census-day" noninterview rate for the P sample

Table 2. Summary of the ICM (PES) Sample: Clusters, Housing Units, and People

	Sacramento	Menominee	South Carolina
No. of Clusters	390	21	674
No. of P-sample HUs ¹	16,419	794	17,677
Interviewed	14,322	409	14,972
Nonint'v'd	765	7	822
Vacant or deleted HUs	1,332	378	1,883
No. of P-sample people ¹	35,509	1,233	35,018

¹ See section 3.3 for a definition of the P sample.

is defined as the number of noninterviewed housing units—based on the Census-day status of the housing unit—divided by the number of P-sample housing units. These rates were 4.7% in Sacramento, 0.9% in Menominee, and 4.7% in South Carolina.

To evaluate the procedures for households that moved into P-sample housing units *after* Census day, we define "interview day" noninterview rates. The numerator here is the number of noninterviewed units based on the status of the housing unit *on the day of the ICM interview*. These rates were 2.0% in Sacramento, 0.1% on Menominee, and 1.6% South Carolina.

3.3 Dual System Estimation in the ICM (PES)

Estimation in the ICM entailed three major steps—dual system estimation, post-stratification, and raking. A brief description of each is given in the following subsections. Dual system estimation (DSE) is based on capture-recapture methodology. Consider only those housing units contained in the sample block clusters selected for the ICM. Within these clusters we define the units enumerated in the initial phase as the E sample, and those enumerated in the ICM as the P sample. After the ICM field work was completed, the Census Bureau tried to match all P-sample persons to E-sample persons. Each person is classified in one of three ways, according to whether he or she is matched (included in both enumerations), included in only the census enumeration (only in the E sample), or included in only the ICM enumeration (only in the P sample). Table 3 demonstrates the possibilities.

Table 3. Enumeration in the Census and in the ICM

ICM Enumeration (P Sample)	Census Enumeration (E Sample)		Total
	In	Out	
In	N_{11}	N_{12}	N_{1+}
Out	N_{21}	(N_{22})	(N_{2+})
Total	N_{+1}	(N_{+2})	(N_{++})

Note that N_{11} , N_{12} , and N_{21} , and thus also the marginals N_{1+} and N_{+1} , are observed after we match the P and E samples. The cell entry N_{22} and all sums derived using N_{22} are unobserved (and placed in parentheses). If we assume that the two enumerations are conducted independently, then the interior frequency N_{11} / N_{++} should be approximately equal to the product of the marginal frequencies $(N_{+1} / N_{++}) \times (N_{1+} / N_{++})$. In that case, we can estimate the unknown total as $N_{++} = N_{+1} \times N_{1+} / N_{11}$. Note that, for any person, the probabilities of enumeration in the initial phase and in the ICM need not be the same. However, the enumerations must be conducted independently.

To this point, the numbers in Table 3 refer to counts or estimates in the ICM sample block clusters. But because the ICM clusters are a representative sample of the appropriate site, we apply the sample weights to get an estimate of the total population of the site. In the DSE model above, N_{1+} represents the number of people enumerated in the ICM (the number of P-sample people). N_{11} is the number of people enumerated in the initial phase and in the ICM, that is, the number of matches between the P and E samples. Then the ratio of weighted totals N_{1+} / N_{11} is the inverse of the weighted match probability among P-sample people.

The weighted estimate N_{+1} is the number of distinct and identifiable census persons (E-sample people), also called the official census count. The official count, however, includes imputed persons and people who are erroneously enumerated. People are imputed when a census enumerator confirms that a certain number of people live at an eligible address, but insufficient additional information can be gathered. Erroneous enumerations include people who should not be counted at that address, for example, because they should be counted elsewhere, such as in a college dormitory, or because their residence is actually at a different address. To correct for these situations, in place of N_{+1} in the formula above, the Bureau subtracts the number of whole-person imputations and multiplies by the portion of correct enumerations (1 minus the rate of

erroneous enumerations, as estimated in the E sample).

In Table 4, for matches, imputed persons, and erroneous enumerations, we display the aggregate rates as measured across each site.

Table 4. Rates Used in the Dual System Estimate

	Sacramento	Menominee	South Carolina
Rate of matches	78.1%	82.9%	74.1%
Rate of imputed persons	7.4%	7.6%	5.6%
Rate of erroneous enumerations ¹	10.5%	9.7%	13.7%

¹ Duplicates, geocoding errors, fictitious persons, illegible names, etc.

It should be noted, however, that the rates are applied at more detailed levels in the actual estimator. In the Dress Rehearsal, these rates were estimated at the site level from information gathered *within the ICM clusters*. For example, follow-up operations were used to determine erroneous enumerations by identifying duplicates, geocoding errors, fictitious persons, and illegible names. Other adjustments are made to the estimator above to account for specific aspects of the ICM operations. For example, different procedures must be applied in the case of households where the census-day residents moved out before the ICM enumeration. This allows us to obtain a more precise estimate of the number of P-sample people and matches in the ICM.

Because capture probabilities are not equal for all members of the population, we try to partition the population into groups (post-strata) such that coverage probabilities are similar for all members in a post-stratum but different in different post-strata. The dual system estimates are then calculated separately in each post-stratum. In the Dress Rehearsal we formed post-strata based on combinations of tenure, race, ethnicity, age, and sex—all done separately within each of the three sites.

For a given post-stratum, the coverage factor is defined as the ratio of the dual system estimate divided by the census count. This coverage factor allows us to compute small-area estimates at the block level using an approach called synthetic estimation. For people satisfying the characteristics of the post-stratum, the block-level estimate (for these people) is obtained by multiplying the corresponding census count by the post-stratum coverage factor. A controlled rounding procedure is then applied to obtain integer person

estimates. For information on these and other more intricate aspects of the dual system estimator as applied to the 1990 Post Enumeration Survey, see Hogan (1993).

3.4 Post-strata and Raking

As mentioned in section 3.3, dual system estimates were to be computed for 84 predefined groups, or post-strata, as given by cross-classifying the following:

Race-ethnicity (6 groups): Hispanic; and non-Hispanic groups of the races black, Asian, American Indian, Hawaiian or Pacific Islander, and white and all others.

Age-sex (7 groups): male or female, 0 - 17; male, 18 - 29; female, 18 - 29; male, 30 - 49; female, 30 - 49; male, 50 and over; female, 50 and over.

Tenure (2 groups): owner, renter.

These post-strata were required to have a minimum population size; if not, groups were collapsed according to predefined criteria. Because of differing racial and ethnic compositions in the dress rehearsal sites, some post-strata were collapsed in each of the three sites. (See Memorandum [3].) However, the two tenure groups—owner and renter—were never collapsed together.

In Sacramento, American Indians, Hawaiian and Pacific Islanders, and Hispanics were placed in one group, but all age-sex and tenure categories were retained throughout the site. This produced 56 (rather than 84) final post-strata in Sacramento. In South Carolina, whites and others formed one race-ethnicity group, but the other five were collapsed into a second group. Retaining all age-sex and tenure categories, the site used 28 post-strata for estimation. The post-stratification was a little more complex in Menominee. There, blacks, Asians, American Indians, and Hawaiian and Pacific Islanders were combined. For this group we retained the age-sex and tenure breakdowns. For the white-and-other group, we collapsed the first four age-sex categories above; the other three stood on their own. For the Hispanic group, we collapsed all age-sex categories, and ended with two post-strata—one each for owners and renters. This produced 24 post-strata in Menominee.

To reduce the variance of the dual system estimates (DSEs), we employed raking procedures using several steps. (1) We assigned the race-ethnicity \times age-sex groups (after collapsing groups as necessary) to the rows of the raking matrix, and the two tenure groups to the columns. (2) The internal cells of the raking matrix were filled with the DSEs corresponding to the row and column categories. (3) The marginals were then obtained by summing the DSEs across rows and down columns of the

matrix. (4) The initial interior cells were then replaced with the *unadjusted population counts*. (5) The interior cells were raked to the marginal totals until convergence. For more details on the post-stratification and on the effect of the raking procedure, see Schindler (1999).

3.5 The Summary of Estimates and Undercounts

In Table 5 the apportionment count for each site is given and divided into its several components. These components include the use of sampling for nonresponse follow-up and UAA vacant returns and multiplicity estimation in Sacramento, and the adjustment for the net undercount in Sacramento and Menominee.

Table 5. Census Counts and Their Components

Components	Sacramento	Menominee	South Carolina
Actual count for GQs ¹ and SBE	7,066	45	33,524
Actual count for NonGQ persons	334,952	4,490	612,962
Actual count for unclassified HUs	7,143	60	15,654
Added sampling for NRFU	24,930	NA ²	NA
Added UAA vacant	2,409	NA	NA
Added SBE	1241	NA	NA
Added ICM	25,572	143	NA ³
Total Count for Apportionment	403,313	4,738	662,140

¹ GQs: group quarters; SBE: service-based enumeration.

² Not applicable.

³ Although no persons were added to the apportionment count in South Carolina, the PES estimated a net undercount of 65,108 people.

The estimated net undercount for a group or geographic area is defined as the difference between the group's census counts after (adjusted) and before (unadjusted) applying the ICM (PES) coverage factors. The net undercount rate is simply this difference divided by the adjusted count. Over the entire site, including people enumerated in group quarters and services such as

shelters and soup kitchens, the net undercount rates (with standard errors in parentheses) were estimated as 6.3% (1.1%) in Sacramento, 3.0% (1.8%) in Menominee, and 9.0% (1.6%) in South Carolina.

Finally, Table 6 provides estimated net undercount rates for the Dress Rehearsal computed for specific race-ethnicity groups retained in the post-stratification for that site (Memos. [2], [3], and [4]). It should be noted that the rates in Table 6 are based only on the housing unit population; group quarters and people enumerated at services are not included in the rate's denominator.

Table 6. Net Undercount Rates for Race-Ethnicity Groups (Housing Unit Population Only) ¹

2000 Census Dress Rehearsal	Net Undercount Rate (SEs in parentheses)
Sacramento, CA	
Non-Hispanic White ²	4.7% (1.2%)
Non-Hispanic Black	8.7% (2.4%)
Non-Hispanic Asian	6.0% (1.9%)
Non-Hispanic American Indian	8.5% (1.7%)
Non-Hispanic Hawaiian	8.0% (1.6%)
Hispanic	8.3% (1.5%)
Total	6.5% (1.1%)
Menominee	
Non-Hispanic Amer. Indian on Reservation	4.1% (2.0%)
Total	3.0% (1.8%)
South Carolina	
Non-Hispanic White ²	6.3% (1.3%)
Hispanic, Black, Asian, Am. Ind., Hawaiian	13.2% (2.6%)
Total	9.4% (1.6%)

¹ These undercount rates do not include people in Group Quarters and Service Based Enumerations in the denominator.

² This group includes Non-Hispanics of all races other than Black, American Indian, Asian, and Hawaiian and Pacific Islanders.

References

Hogan, H. (1993). "The 1990 Post-Enumeration Survey: Operations and Results," *Journal of the American Statistical Association*, 88, 1047-1060.

National Research Council (1995). *Modernizing the U.S. Census*, Panel on Census Requirements in the Year 2000 and Beyond, National Academy Press, Washington, DC.

Schindler, E. (1999). "Iterative Proportional Fitting in the Census 2000 Dress Rehearsal," *Proceedings of the Section on Survey Research Methods, American Statistical Association* (to appear).

Bureau of the Census Internal Memoranda

[1] Memorandum from Donna L. Kostanich to Dennis W. Stoudt, "Census 2000 Dress Rehearsal Initial Phase Estimation: Approval and Summary of Results," February 16, 1999, DSSD Census 2000 Dress Rehearsal Memorandum Series, #A-74.

[2] Memorandum from Rajendra P. Singh to Howard Hogan, "Some Results from the Census 2000 Dress Rehearsal," February 26, 1999, DSSD Census 2000 Dress Rehearsal Memorandum Series, #A-76.

[3] Memorandum from Donna L. Kostanich to Dennis W. Stoudt, "Approval of 2000 Dress Rehearsal Site Level Estimation," March 15, 1999, DSSD Census 2000 Dress Rehearsal Memorandum Series, #A-80.

[4] Memorandum from Donna L. Kostanich to Howard Hogan, "Preliminary Investigation of South Carolina Post Enumeration Survey Results," April 5, 1999, DSSD Census 2000 Dress Rehearsal Memorandum Series, #A-83.

Acknowledgment

The authors thank John Bushery for his extensive review and invaluable comments.

**This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.*

RESPONDENT UNDERSTANDING OF DECENNIAL CENSUS RESIDENCE INSTRUCTIONS: THREE PRETESTING METHODS, THREE RESULTS

Donna Eisenhower, National Opinion Research Center; Karen Mills, U.S. Bureau of the Census;
Eleanor Gerber, U.S. Bureau of the Census; and Lisa Lee, NORC

Karen Mills, Population Division, U.S. Bureau of the Census, Washington, DC 20233

Key Words: Household Enumeration, Pretesting
Methods, Vignettes

1.0 Introduction

An accurate and complete count of persons is essential for the decennial census. Most of the household enumeration is based on self-administered questionnaires returned by mail. The respondent is provided a list of residence instructions on the front page of the form, which guides the respondent in determining whom to include or exclude on the form. The quality of the overall census count depends, in a significant sense, on how accurately the respondent reads and interprets these instructions for his or her household enumeration, or on how well the respondent's own natural decision rules for inclusion and exclusion coincide with those of the census residence instructions.

The purpose of this research was to compare and understand the nature of the response accuracy between two sets of residence instructions. The longer set of instructions had been used on the 1990 census questionnaires, and the shorter set, on the Census 2000 Dress Rehearsal questionnaires. Each instruction set had "include" and "do not include" headings, with examples of situations listed underneath. The longer set had 13 instructions (8 "include" and 5 "do not include" instructions). The shorter set had 7 instructions (3 "include" and 4 "do not include" instructions).

Earlier research had suggested that it might be effective to use a shorter set of instructions if the omitted instructions coincided with the respondent's natural decision rules or intuition for including or excluding people on the questionnaire. Therefore, some of the rules that were repeated on both the "include" and "do not include" lists for the longer set were eliminated for

the shorter set. For example, on the longer set, "College students who stay here while attending college" was under the "include" list, and "College students who live somewhere else while attending college" was under the "do not include" list. Both sets of instructions rest on the underlying concept of "usual residence," that is, the place where the person lives and sleeps most of the time.

The response accuracy of the shorter and longer versions of the instructions was tested using three different pretesting methods, namely: (1) group administrations of a vignette methodology, (2) focused group debriefings, and (3) cognitive interviews. This paper discusses the different kinds of information and insight provided by these three pretesting methods. The findings of this effort are discussed within the context of the differential use of pretesting methods. The singular dimension of the residence instructions compared to a lengthy questionnaire with a variety of questions provides a good opportunity for such a methodological assessment.

2.0 Methodology

The relative accuracy of responses using the two sets of residence instructions was assessed using a qualitative methodology that consisted of the following:

- *Group administrations of vignettes*, in which people in two groups (Group 1 and Group 2) were asked to answer questions about 17 vignettes. Vignettes, or brief narratives, present hypothetical people in ambiguous living situations, suggesting that there is more than one place where a person could be counted. Prior to the administration of the vignettes, Group 1 was asked to complete a household questionnaire containing the shorter set of residence instructions; Group 2 was asked to complete the same questionnaire with the longer set. The list of vignettes is shown in section 6.0 at

This paper reports the results of research and analysis undertaken by Census Bureau staff and researchers at the National Opinion Research Center. It has undergone a more limited review than official Census Bureau publications. Research results and conclusions expressed are those of the authors and have not been endorsed by the Census Bureau. This report is released to inform interested parties of research and to encourage discussion.

the end of this paper.

- **Focused group debriefings**, in which members of both groups were asked to think about and evaluate different aspects of the residence instructions and vignettes using a series of questions and an open-ended discussion.
- **Cognitive Interviews**, in which five participants in each group were asked retrospectively to "think aloud" about how they reached answers to selected vignettes.

2.1 Participant Mix of the Groups Administered the Vignettes

Fifty-four people were assigned to one of two groups (28 and 26 persons for Group 1 and Group 2, respectively). The desired mix of participants was driven most by educational level, with 16 recruits having less than a high school education, 17 having only a high school education or GED, and 9 having only some college, and 12 having a college or postgraduate degree. The groups also had broad racial and ethnic representation as well as a wide representation of households by size and special circumstances, for example, persons away at college, persons in institutions such as nursing homes, persons with extended family living with them, and persons with roomers or foster children. A similar representation of characteristics was achieved for each of the two groups.

2.2 Group Administrations of the Vignettes

The vignettes provided an objective measure of accuracy to assess differences between the two groups. Most of the vignettes were based on previous work reported by Gerber et al. (1996); however, some were created specifically for this exercise to cover situations in the longer list of residence instructions. The vignettes were developed from ethnographic sources, including ethnographic interviews on residence (Gerber, 1990), a Living Situation Survey pretest report (Sweet, 1994), ethnographies commissioned by the Census Bureau to examine the behavioral causes of the undercount (de la Puente, 1993), and a report on the experience of Hispanics in the 1990 census (Kissam et al., 1993). This ethnographic basis for the vignettes ensured that they were perceived as "real" situations. As further reported by Gerber, the vignettes had neutral vocabulary and avoided technical census residence

terms.

The facilitator read the vignettes aloud and also gave the participants a written version to read on their own. They recorded their answers after each vignette. There was no time pressure, and participants readily answered the questions on the vignettes without hesitation or complaints.

2.3 Focused Group Debriefings

Group 1 and Group 2 were each divided into three subgroups for the debriefings. The five who were to be cognitive interview participants formed one subgroup, and the rest of the participants were assigned to one of two other subgroups. All three subgroups were asked eight closed-ended, precoded debriefing questions about the vignettes and residence instructions. The facilitator read the debriefing questions, and participants also were given time to read them before entering their answers.

2.4 Cognitive Interviews

Ten participants were selected for the cognitive interviews, five from Group 1 and five from Group 2. The cognitive interview provided an opportunity to explore what was going through the respondents' minds as they answered the vignettes. The retrospective "think aloud" technique was used, whereby the respondent was asked to reread the vignette and think aloud about what was going through his or her mind in answering the questions. The vignettes of focus were those for which the respondent had given an incorrect answer based on census-defined inclusion and exclusion criteria. The vignette answer sheet was scored prior to the interview, and the respondents were generally aware if they had a different answer from what the census designated as correct. A relaxed atmosphere was maintained, and interest was focused on understanding *how* respondents came up with their answers, not on the answers themselves.

3.0 Findings by Pretesting Method

3.1 Findings From the Vignette Administrations

The analysis of the vignette administrations showed a clear pattern: Group 2 (with the longer set of residence instructions) performed better overall than Group 1

(with the shorter set) in correctly answering the vignettes. (See section 6.0 for vignette list.) In addition, except for V10 (Kathy's roommate), Group 2 scored higher on the individual vignettes. Three scores reached significance (.05 level or better), as follows:

- The overall difference score: $p < .002$, with 65.4% of Group 1 answering correctly, compared with 78.6% of Group 2
- V7- Jack: $p < .004$, with 42.3% of Group 1 answering correctly, compared with 82.1% of Group 2
- V9- Dave/Johnsons: $p < .001$, with 69.2% of Group 1 answering correctly compared with 100% of Group 2

V1- Maria approached significance at $p < .064$, with 34.6% of Group 1 answering correctly, compared with 60.7% of Group 2.

Fisher's Exact Test (2-tail) was used for the individual vignettes. It tells if the result for a given vignette could have happened by chance. It is preferred to the chi-square test to assess significance levels for small sample sizes. The t-test was used for the overall score. It not only addresses what occurred by chance, but also aggregates across all vignettes and respondents. Therefore, if Group 2 consistently did a little better on most of the items, one would get a significant result using the t-test, as occurred in this case. Small differences can add up to a significant difference in the t-test.

Some variations between the shorter and longer instruction lists may have accounted for the differences in accuracy. For example, the longer instructions:

- Were more prominent on the page because of the form design and greater length of the form itself.
- Had more complete and fewer truncated sentences, which may have eased comprehension for those who read the instructions.
- Had greater redundancy, which provided more opportunity to encode the information. For example, Armed Forces personnel who were listed under both the "include" and "do not include" headings.
- Had less ambiguity by providing more explicit instructions, which required less respondent interpretation or judgment.
- Had more noun-first phrasing (that is, college

student, newborn baby), thereby facilitating skimming of the instructions.

- Had features which better overcame respondents' natural tendency to include a person with the family unit, rather than where he/she lived or slept most of the time.
- Had different design features and layout than the shorter list.

3.2 Findings From the Group Debriefings

The debriefing participants were asked to discuss four open-ended questions. The dynamic of the group context brought out stronger, more extreme opinions and sometimes more tangential opinions (such as about the census undercount and racial issues) than the four questions. While a complete list of comments is not possible here, some examples are as follows:

Examples of comments from groups with shorter version of residence instructions

- Why can't you say what you mean and be more direct?
- If you would emphasize the date of the census, then logic would hold.
- A person can read the instructions and still not understand them.
- If a person pays rent, they should be counted.
- The instructions are boring.
- A person might be pressed for time; there is no time to read them at home.

Examples of comments from groups with longer version of residence instructions

- It would help to have larger, bolder print.
- There is some inconsistency in the rules.
- To be at one place for four days and the other for three is an "iffy" situation.
- People follow their own logic.
- I would skim the instructions.
- I do not need them [the Census Bureau] to tell me who to include in my household.

3.3 Findings From the Cognitive Interviews

The results of the ten cognitive interviews provided insight into "how" respondents came to enter the answers they did regarding the vignettes. While the insight from these interviews showed little between-

group differences, the results did show some overall respondent patterns for providing answers to the vignettes; which, in turn, uncovered patterns or "rules in effect" used to decide whether to include or exclude a person in a household.

While most participants did not carefully read the actual instructions, most participants did make consistent decisions based on these "rules in effect." In other words, many people made up rules and then applied them consistently. The five most prevalent "rules in effect" comments made from participants in both Groups 1 and 2 are the following:

- A person should be listed with his or her family (especially spouse and children). Most participants made decisions based on this premise. A typical comment was "they should be listed where they live, you know, with their family; that is where they really reside." Therefore, relatively high numbers missed the first two vignettes (V1-Maria and V2-Craig).
- You should follow the same rules as for doing your taxes. If you list a person on your tax form, then you should list that person on your census form. "Tax dependency should be the guiding principle." "If you pay for them, you should get credit for them." One participant also mentioned listing the person where he or she has a driver's license. Both the tax form and the driver's license follow a logic related to a "legal" or standard way of doing things.
- Exclude people living in temporary situations. Therefore, Carolyn's mom (V4), placed in a nursing home on a "trial basis," was felt to belong with the daughter, and one participant felt the Smiths in V6 should not be included with the Haywards since they were only staying there until their house was complete.
- You should include the person where he or she was as of April 18 (Census Day for the Dress Rehearsal), regardless of circumstance. For example, many participants missing V17 felt that the Wongs should be listed in Florida since they spend a significant amount of time there, and they were there on Census Day. One person asked, "What if the Wongs do not get a form in Ohio?"
- If a person is self-sufficient, then he or she should get his or her own census form. Therefore, Dave, the roomer in V9, "has his own mailbox and should have his own form since he is paying

rent...he is not a freeloader." Also, Sergeant Kathy in V8 gets her own paycheck, so she should get her own form. This type of comment was made from only one participant from each group. It also carries a sense of legality with it.

4.0 The Three Pretesting Methods: Results Compared

The three pretesting methodologies--vignettes, group debriefings, and cognitive interviews--provided valuable but different insights into the differences between the two sets of residence instructions and the use of the residence instructions in general.

The *vignette method* provided perhaps the most accurate and objective basis of comparison, but their use had some positive and some less positive features as a methodology for assessing differences. The vignettes provided an objective accuracy score and enabled a more quantifiable comparison. They also provided each group member with an array of situations, which then permitted a more complete appraisal of his/her decision-making under varying household situations. The vignettes were also concrete and understandable to those with less than a high school education.

Nevertheless, the vignettes are somewhat artificial, and the circumstances under which the vignettes were completed were somewhat different than those under which the respondent would be answering in the census itself. Most respondents said that when they read the residence instructions, they read them more carefully when answering the vignettes than they would have if they had read them on their own form, particularly if completing it at home. This suggests that the situation might be worse for those completing the form for their own households, and that it would require a far greater number of respondents to capture and assess the range of situations presented in the vignettes. Therefore, it is important not to ignore the findings from the vignettes but rather to understand them.

Finally, the vignette methodology enables one to reach the outcome state "reporting" of the information processing model, while other pretesting methods often fall short of this. In this case, the accuracy of the respondent's judgment to place a person in a vignette

in a given household could be assessed. Tourangeau (1984) proposed that a respondent's answer to a question is the result of four stages, namely: comprehension of the question, retrieval of relevant materials from memory, the processing of this information to form a judgment, and the selection of the appropriate response alternative. Eisenhower et al. (1991) added encoding to the model, whereby information of the event is recorded into the mind in the first place - information may not be recorded, it may be incomplete or it may be distorted. The vignette methodology as a pretesting method permits assessment of the final outcome accurate or inaccurate reporting.

The *focused group debriefing method* provided greater insight into the vignette findings. It provided more information on motivating factors to read or ignore aspects of the instructions. It also provided the clearest indication of the strength of opinions. In the context of the group dynamic, more opinions about the instructions, vignettes, forms, and the census itself were shared. However, the discussion part of the debriefing produced fewer precise, more subjective results and, therefore, required greater care in interpreting. Precoded responses to eight questions collected from group participants prior to the focused discussion provided more precise measures of opinions.

Finally, the *cognitive interview method* provided a different kind of insight that would not have otherwise been uncovered. The cognitive interviews for the vignettes were retrospective "think alouds" with only general questions and a minimum of probing. Yet as soon as the respondents began to "think aloud," it was obvious that they were creating their own rules for deciding whether to include or exclude the person in the vignette at a given household. These "rules in effect" were given without hesitation and yet were not illuminated by any other methodology. The findings from the cognitive interviews did not point to differences between groups, but rather to an in-depth understanding of decision-making strategies used by participants from both groups regardless of the version of the instructions. Some of these cognitive findings suggested the following:

- Some respondents want to include an individual with his or her family, especially with a spouse and children, even if the individual spends the

greater part of the time working elsewhere. They have an affiliation orientation.

- Some respondents have a "legal" sense that results in wanting to include persons on the census form if they are included on the tax form. One individual also was inclined to include the person in the place where the person had a driver's license.
- Finally, other respondents want a simple rule to follow. Two of these simple rules included: 1) all those living here most of the year; and 2) all those living here as of the Census Day.

The cognitive interviews provided more information about how respondents encoded and comprehended the information from the instruction lists. In particular, the interviewer informed researchers of how "judgments" were made to include or to exclude a given person in a given household.

This paper has discussed the research on respondent understanding of the census residence instruction lists to illustrate how different pretesting methods provide different aspects for diagnosing measurement error. In this case, the group administration of the vignette methodology, the focused group debriefings, and the cognitive interviews provided valued but different kinds of results. The results from just one pretesting methodology would have provided only part of the answer and could have been misleading. However, the use of the results of the three methods provided a better basis for minimizing errors in respondent use and in interpretation of the residence instructions in the future. This, in turn, potentially could minimize measurement error and lead to a more accurate count for future decennial censuses.

5.0 References

de la Puente, Manuel (1993). "Why Are People Missed or Erroneously Included by the Census: A Summary of Findings From Ethnographic Coverage Reports." Paper presented at the 1993 U.S. Bureau of the Census Research Conference on Undercounted Ethnic Populations.

Eisenhower, Donna, and L. Lee (1998). "Research on Respondent Understanding of the Census Residence Instruction Lists." National Opinion Research Center. Submitted to the U.S. Bureau of the Census under Task Order No. 46-YABC-8-00003.

Eisenhower, Donna, N. Mathiowetz, and D. Morganstein (1991). "Recall Error: Sources and Bias Reduction Technologies" in Biemer, Paul, et al. (eds.), *Measurement of Survey Error in Surveys*. Wiley, New York.

Gerber, Eleanor (1990). "Calculating Residence: a Cognitive Approach to Household Membership Among Low-Income Blacks." Report prepared for the Center for Survey Methods Research, U.S. Bureau of the Census.

Gerber, Eleanor, T. Wellens, and C. Keeley (1996). "Who Lives Here?: The Use of Vignettes in Household Roster Research." Joint Proceedings of the American Statistical Association. Paper presented to the American Association for Public Opinion Research.

Kissam, Edward, E. Herrera, and J. Nakamoto (1993). *Hispanic Response to Census Enumeration: Focus and Procedures*. Submitted to the U.S. Bureau of the Census under Task Order No. 46-4ABC-2-66027.

Sweet, Elizabeth (1994). "Roster Research Results From the Living Situation Survey." Paper prepared for the American Statistical Association, Section on Survey Research Methods.

Tourangeau, Roger (1984). "Cognitive Science and Survey Methods" in T. Jabine, J. Tanur, and R. Tourangeau (eds.), *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, Washington, DC, 73-100.

6.0 List of Vignettes

[Percent correct for Group 1 (shorter instructions) and Group 2 (longer instructions) and p values shown in parentheses]

1. Maria is a live-in housekeeper for the Smiths during the week, but spends weekends with her husband and children at their apartment. Where should Maria be listed on a census form? *Correct answer: with the Smiths*
(Group 1- 34.6%; Group 2- 60.7%; $p < .064$)
2. Craig and his wife have a house in Pennsylvania. Craig's job is in Washington, DC, so he stays with his mom in DC, Monday through Thursday of the week. Where should Craig be listed on a census form?
Correct answer: Washington, DC
(Group 1- 23.1%; Group 2- 46.4%; $p < .092$)
3. Sergeant Jim is stationed in Alaska while his family has stayed behind in Maryland. Should Jim's wife put him on her census form? *Correct answer: no*
(Group 1- 53.8%; Group 2- 67.9%; $p < .403$)
4. Carolyn's mom normally lives with her; however, on [date before Census Day in the census month] she has placed her mom in a nursing home on a trial basis for the next 3 months. Should Carolyn put her mom on her census form? *Correct answer: no*
(Group 1- 30.8%; Group 2- 46.4%; $p < .275$)
5. Mary's daughter Alice has been away at college and has 3 more years until graduation. Should Mary put her daughter on her census form? *Correct answer: no*
(Group 1- 57.7; Group 2- 75.0%; $p < .250$)
6. The Haywards are sharing their apartment with the Smiths until the Smiths' new house is complete. Should the Haywards list the Smiths on their census form?
Correct answer: yes
(Group 1- 46.2%; Group 2- 60.7%; $p < .413$)
7. Jack does not have a place of his own, but stays part of the time at his mother's house and part of the time at his sister's house. On [census day], he was at his sister's. Where should Jack be listed on a census form?
Correct answer: his sister's house
(Group 1- 42.3%; Group 2- 82.1%; $p < .004$)
8. Sergeant Kathy is stationed at a base near her mother's house, and stays with her mother. Should her mother put Kathy on her census form? *Correct answer: yes*
(Group 1- 80.8%; Group 2- 89.3%; $p < .460$)
9. Dave rents a room at the Johnsons' house. Should the Johnsons list Dave on their census form?
Correct answer: yes
(Group 1- 69.2%; Group 2- 100.0%; $p < .001$)
10. Kathy's roommate moved in on [4 days before census day]. Should Kathy list her roommate on her census form? *Correct answer: yes*
(Group 1- 88.5%; Group 2- 78.6%; $p < .470$)
11. Mary stayed with her friend Sue for the first 2 weeks in [census month] and then returned to her apartment in Seattle. Should Sue list Mary on her census form?
Correct answer: no
(Group 1- 92.3%; Group 2- 92.9%; $p < 1.000$)
12. Romero lives at home with his parents while attending college. Should they list Romero on their census form?
Correct answer: yes
(Group 1- 92.3%; Group 2- 96.4%; $p < .604$)
13. Ned is an eighth grader at a boarding school in another city but comes home for holidays and summers. Should Ned's parents put him on their census form?
Correct answer: yes
(Group 1- 69.2%; Group 2- 82.1%; $p < .346$)
14. Tonya and her newborn baby are still in the hospital. Should Tonya's husband list the baby on the census form? *Correct answer: yes*
(Group 1- 96.2%; Group 2- 100.0%; $p < .481$)
15. Joan and Tommy's parents are divorced. Their father keeps a room for them at his apartment, and they are

named on his lease. They spend weekends with him. Should Joan and Tommy's father list them on his census form? *Correct answer: no*

(Group 1- 73.1%; Group 2- 85.7%; $p < .320$)

16. Esther has a foster child assigned to her by the city. The child has been there for 3 months, and Esther doesn't know how long the child will stay there. Should she list the foster child on her census form?

Correct answer: yes

(Group 1- 92.3%; Group 2- 96.4%; $p < .604$)

17. The Wongs have a vacation home in Florida, where they stay from January 5 to May 1. They then return to their house in Ohio. They receive a census form at their Florida house. Should they list their names on this form? *Correct answer: no*

(Group 1- 69.2%; Group 2- 75.0%; $p < .764$)

SELECTING VARIABLES FOR POSTSTRATIFICATION AND RAKING

Golam M. Farooque, Inez I. Chen, Bureau of the Census
Golam M. Farooque, Bureau of the Census, Washington, DC 20233

KEY WORDS: Logistic regression; Interaction terms; Raking dimensions; Post Enumeration Survey.

ABSTRACT: This article applies logistic regression models to the 1990 Post Enumeration Survey (PES) data for California and determines the important variables to form alternative poststratifications and raking matrices. The person level indicator variable for capture in the census is used as the dependent variable. This paper finds that age/sex, race/Hispanic origin, tenure, household composition, and urbanicity variables are the most important variables for forming alternative poststratifications and raking matrices. The first order interaction terms of significant independent variables are found insignificant when they are input to the logistic regression models with their main effects.

1. Introduction

For Census 2000, a major goal of the Census Bureau is to reduce the undercount, especially the differential undercount in different segments of the population. Generally, undercounts tend to be higher for minorities, especially, for Blacks, Hispanics, Asian, and nearly all nonowners (Hogan 1993, Robinson et al. 1993). The Bureau has been using Dual System Estimation (DSE) (capture/recapture technique) with post-stratification to produce Census counts at various geographical levels in order to correct for coverage errors. The DSE assumes that the probabilities of enumeration are the same for all members of the population. The past research showed that the probability of being enumerated in the census varies by race, age, sex, tenure, and geographical areas, hence, the homogeneous probabilities' requirement for DSE is not met. A considerable number of studies have been conducted to provide improved estimates of persons missed by both initial enumeration and the PES enumeration (Hogan 1993, Alho, et al. 1993).

The authors are mathematical statisticians in the Decennial Statistical Studies Division. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau Publications. This report is released to inform interested parties of research and to encourage discussion. The views expressed are attributed to the authors and do not necessarily reflect those of the Census Bureau.

Yet, a residual heterogeneity remains which may cause correlation bias in the DSE. Alho et al. (1993) and Murly et al. (1997) used logistic regression analysis to select variables to reduce heterogeneity. Currently, the Bureau is conducting research on how to identify a set of variables that will be used to form poststrata for the year 2000.

This paper conducts a preliminary research to identify the important variables, from a set of variables to form alternative poststratifications and raking matrices. It is expected that this research will aid the 2000 poststratification research. By applying the logistic regression models, this paper identifies the important variables from the 1990 PES data for California. This research is similar to Haines and Hill (1998) study. However, it incorporated the two-way interaction terms in addition to main effects in the logistic regression models. To determine the statistical significance of the variables, it used global tests and deviance tests in addition to Wald tests.

A literature review is presented in Section 2. Section 3 presents the methodology which includes definitions of variables, logistic regression model, model inferences, and variable selection techniques. The results of logistic regression models are presented in section 4. Section 5 concludes the paper.

2. Literature Review

The Census Bureau has been using the logistic regression modeling as an analytical tool to analyze the effects of some geographic and demographic variables in a multi-variate modeling of a dichotomous dependent variable (Alho et al. 1993, Mulry et al. 1997, Haines and Hill 1998). Haines and Hill applied logistic regression models on race/origin, age/sex, tenure, urbanicity, household composition (HH comp), household marital status (HH marital status), percent nonowners, mail response rates (MRR), percent minority, vacancy rates, household size, and relationship. Using the 1990 PES data for California, they identified that race/origin, age/sex, tenure, urbanicity, MRR, and relationship are important variables for raking matrices.

Hogan (1993) anticipated that many of the 1,392 poststrata adjustment factors would have very high variances and suggested to form fewer poststrata for the 1990 PES data. The adjustment factors were calculated by dividing the estimated true population by the census

count. Assuming homogeneity exists in fewer poststrata, he regressed observed adjustment factors by poststratum on indicator variables such as race, Hispanic origin, age, tenure, census division, place/size category, and some two-factor interaction terms. He forced race, age, and tenure indicator variables in the model and the other variables were selected using their predictive power.

Alho et al. (1993) applied the conditional logistic regression models to the 1990 PES data for minorities to estimate the probability of being captured in a census. The independent variables were both continuous and categorical. They used age, female, Black, Hispanic origin, renter, household size, related, married, % renter, %Black, %Hispanic, %multi-unit, %vacant, four census regions, metropolitan, and some interaction terms in their models. The dependent variable consists of persons who are captured only in the E-sample, only in the P-sample, and in both E and P-samples. They excluded the unresolved cases from the data. Their results reveal that household size, Black, Hispanic, renter and relationship are significant variables, in at least one of the census regions.

Mulry et al. (1997) examined the heterogeneity of census coverage error for small areas. Using a similar set of variables such as used by Alho et al., Mulry et al. built four separate logistics regression models for minorities, non-minorities, owners, and renters.

Using the 1994 National Assessment for Educational Progress (NAEP) data, Wallace and Rust (1996) studied the performance of raking and poststratification. Their estimated regression model consisted of age, race, regions, school types, metro status, and two-way interactions of these variables. Using a SAS regression procedure they developed the initial models. Since SAS procedure assumes simple random sampling, Wallace and Rust further estimated the selected initial models by using WesVarPC software.

3. Methodology

The methodology section comprises several subsections. In section 3.1, the data and variables are discussed. Sections 3.2, 3.3, 3.4, and 3.5 are on logistic regression modeling, odds ratios, model inference, and variable selection techniques, respectively.

3.1 Data and Variables

The study uses the 1990 PES data for California. The PES consisted of E-sample and P-sample. The E-and P-samples are overlapped. That is, both samples are from the same census blocks and housing units within blocks. The P-sample estimates the number of people missed by the original enumeration and the E-samples estimates the

number of enumerations that are erroneous. This paper uses the resolved PES data (i.e., E- and P-samples matches, E-sample non-matches, and P-sample non-matches). Alho et al. (1993) argue that logistic regression procedures yield invalid parameter estimates if they are fitted to unresolved data. An unresolved case may arise due to a fictitious individual or the available information about the person is ambiguous.

The dependent variable in the model is a person-level indicator. It is assigned the value 1 for matched persons between E- and P- samples and non-matched E-sample persons, and 0 for non-matched P-sample persons.

The Bureau has been forming poststrata based on age, sex, race/origin, tenure, census region, and urbanization variables. Historically, it is known that these variables are good predictors of the probability of a person being included in the census.

Following previous research, in particular, Haines and Hill (1998), this paper applies the logistic regression to a set of independent variables- 5 categories of race/Hispanic origin (Non-Hispanic White and Other, Black, Non-Black Hispanic, Asian & Pacific Islander, and American Indians on Reservations), 7 categories of age/sex (under 18, 18-29 Male, 18-29 Female, 30-49 Male, 30-49 Female, 50 + Male, and 50 + Female), 2 categories of each of the following variables: tenure(owner or renter), HH comp (easy to enumerate or hard to enumerate), HH marital status (spousal or non-spousal), relationship(related or not related), urbanicity (urban or non-urban), percent owners (low or high), MRR (low or high), percent minority (low or high), and household size (one or (two & more)). These variables seem to be good predictors of the probability of a person being included in the census.

This paper tested two definitions for household composition and urbanicity variables. Household composition 1 is a household level variable and household composition 2 is a person level variable. Urban 1 consists of all urbanized areas against non-urban areas. Urban 2 contrasts large urban areas to all other areas. The results reported in the paper are based on urban 1 because urban 2 was found insignificant. Hence, urban 1 is referred as urban for the rest of the paper.

In addition to the main effects, the following two-factor interactions of significant main effects are also tested: race/Hispanic*age/sex, race/Hispanic*tenure, race/Hispanic*household composition 2, race/Hispanic*urban, age/sex*tenure, age/sex*household composition 2, age/sex*urban, tenure*household composition 2, tenure*urban, and household composition 2*urban. Interactions with household composition 2 are tested only because it is found more significant than household composition 1.

All independent variables are converted to indicator variables. For variables with more than two categories, the number of indicators equal to number of categories - 1; for variables with two categories, 1 indicator variable is created.

3.2 Logistic Regression Model

The model used in this paper is a binary model. This model is used to determine the probability of an individual with a given set of attributes will make one choice rather than the other.

Due to many assumptions required for the ordinary least square method, it cannot be applied to analyze data with binary dependent variable because (a) the relationship between the dependent variable and the predictors are nonlinear, (b) the error term and the predictors are correlated, and (c) the presence of heteroscedasticity in error term.

This study has chosen the logistic regression model over other binary choice models because it has been used in the Bureau's research. This paper develops the following logistic regression model with the situation where the independent variables are dichotomous.

$$(1) \Pi(Y|D) = e^{\alpha + \beta D} / (1 + e^{\alpha + \beta D})$$

Where, $\Pi(Y|D) = P(Y=1|D)$, the probability that the person is captured in the Census for a given set of characteristics and $1 - \Pi(Y|D) = P(Y=0|D)$, the probability that the person is not captured in the Census for a given set of characteristics. D is a vector of independent indicator variables that defines the characteristics of a person. β is a set of parameters associated with variables, D s, and α is the intercept.

The parameters α and β s are estimated by applying the maximum likelihood estimation technique on equation (1). Estimation proceeds by finding estimates for α and β s that maximizes the likelihood function (2) for the set of values y_1, y_2, \dots, y_n with given probability defined by D_i . n is the sample of independent observations.

$$(2) l(\alpha, \beta) = \prod_{i=1}^n \Pi(D_i)^{y_i} [1 - \Pi(D_i)]^{1-y_i}$$

3.3 Odds and Odds Ratios

For large samples, the parameter estimates tend to follow a normal distribution. By substituting the estimated parameters α and β s in equation (1) one can easily calculate the probability of a person captured in the census. Once the probability of a person being captured is known, then the odds of that person being captured in

the census can be computed as

$O = \Pi(Y|D) / \{1 - \Pi(Y|D)\}$ for a given set of characteristics,

D. The odds ratio is the ratio of the odds.

The logit transformation of $\Pi(Y|D)$ is

$$\log \left[\frac{\Pi(Y|D)}{1 - \Pi(Y|D)} \right]. \quad \text{With the relationship expressed in}$$

equation (1), this transformation can be expressed as an additive function of independent variables such as equation (3).

$$(3) \log O = \alpha + \sum_{i=1}^K \beta_i D_i + \varepsilon$$

Where, $\log O$ is the log odds of being captured in the census given the explanatory variables. K is the number of independent variables included in the model. β_i is the estimated regression coefficient of the i th independent variable. It estimates the change in the log odds of being in categories of interest on the response for a one-unit change in the i th independent variables in the model.

3.4 Model Inferences

The statistical significance of variable(s) is assessed using the Wald-test and deviance test. The Wald test is computed by squaring the standard normal test when $\beta = 0$. It is approximately distributed as chi-square. The deviance test is the likelihood ratio statistic comparing the reduced model to the full model; it is the statistic for testing the hypothesis all parameters that are in the full model but not in reduced model are equal to zero. The global test is conducted to determine the adequacy of the model.

The calculated Wald test statistics, deviance test statistics, and global test statistics are compared with an adjusted critical value of chi-square. The critical values of $\chi^2(\nu)$ are multiplied by a design effect (DE) to reflect the sample design difference between PES and simple random sampling. All models in this paper are estimated by using the SAS PROC LOGISTIC procedure. This procedure assumes that data is produced by simple random sampling. For California, $DE = 20.2$ was used. The DE is the ratio of the variance of the 1990 PES undercount and the variance under simple random sampling. For the 2000 poststratification research, SUDAAN software will be used and SUDAAN accounts for complex sample design.

3.5 Variable Selection Methods

This paper first determines the significance of main effects and then tested the significance of two-factor interactions of significant main effects. There are several

variable selection methods available. However, none of these methods are appropriate for this study because we forced race/Hispanic origin, age/sex, and tenure variables in the model. Historically, these variables were found significant and used to form poststrata. Thus, instead, this paper used the following two variable selection methods to select the important variables. In method 1, age/sex, race/Hispanic origin, and tenure are always forced in the model and the other independent variables tested individually. Haines and Hill (1998) used this method to select the variables.

In method 2, we also forced age/sex, race/Hispanic origin, and tenure variables in the model. However, unlike method 1, in method 2, the variables which were found significant in the previous steps were kept in the model while testing another new variable. A variable is considered significant if the calculated χ^2 exceeds the adjusted critical values of χ^2 at least at 10 percent levels of significance.

4. Results

Table 1 shows the estimated logistic regression results which are obtained by using variable selection method 1. Based on the results of Table 1, we find that household composition 1, household composition 2, urbanicity, and MRR are statistically significant. However, household composition 2 is significant at 1 percent level, household composition 1 and urbanicity variables are significant at 5 percent level, and MRR is marginally significant at 10 percent level. This decision is reached based on the Wald test statistics and deviance tests reported on this Table. As mentioned before, the critical values for these tests are adjusted for a design effect =20.2. Using the Global test, one can also conclude that the model with household composition 2 in presence of age/sex, race/Hispanic origin, and tenure performs better than any other models.

Table 1. Logistic Regression for Method 1 (Bolded Variables were Forced)

Independent Variables	Wald Test	Odds Ratio	Global Test	Deviance Test
1. race, age/sex, tenure	-	-	725.68(11)*	-
2. HH Comp 1	87.41(1)**	1.801	820.16(12)*	94.48(1)**
3. HH Comp 2	174.79(1)*	2.240	916.55(12)*	190.87(1)*
4. HH marital status	25.07(1)	1.467	751.74(12)*	26.06(1)
5. relationship	1.189(1)	0.942	726.87(12)*	1.19(1)
6. urban	91.42(1)**	2.136	805.06(12)*	79.38(1)**
7. % nonowners	4.42(1)	1.116	730.12(12)*	4.44(1)
8. mail response rates	54.10(1)***	0.706	778.18(12)*	52.51(1)
9. % minority	12.65(1)	0.828	738.34(12)*	12.67(1)
10. household size(HHS)	17.36(1)	0.697	741.96(12)*	16.28(1)

Note: On all Tables *, **, and *** mean significant at 1, 5, and 10 percent levels of significance, respectively.

The odds ratios of estimated coefficients of logistic regression models can be used to explain the odds of a person being captured in the census. For example, in Table 1 (model 3), the odds ratio for easy to enumerate persons versus hard to enumerate persons is 2.240. This means that odds of being captured for easy to enumerate persons are about 2.240 times larger than hard to enumerate persons. Similarly, the odds of urban people

(model 6, Table 1) being captured in the census is 2.136 times higher than non-urban people.

Table 2 displays the odds ratios, deviance test, and global test statistics for logistic regression models by using the variable selection method 2. A variable is considered significant if its estimated test statistic is less than a pre-specified significance level of 10 percent. The importance of each other variable is evaluated by keeping the previous significant tested variable(s) in the model.

Since the variable selection method 1 indicates that HH comp 1 and HH comp 2 both are significant, two separate models are estimated: one with HH comp 1 and another with HH comp 2. The results are on Table 2. On this Table, the first set of statistics corresponding to each variable represents the test statistics produced from models with HH comp 1 and the second set in brackets under the first set denotes the test statistics obtained from models with HH comp 2. Like variable selection method 1, this method also shows that HH comp and urbanicity variables are significant. However, HH comp 2 is more significant than HH comp 1.

We also test whether different order of entering the variables in the models makes any difference in results. The findings are similar to those of Tables 1 and 2. Results are available from the authors upon request. Thus based on these findings we conclude that race/origin, age/sex, tenure, HH comp 2 and urbanicity are important variables to form poststratification and alternative raking matrices for California.

Next, we proceed to determine the significant two-way interactions for significant main effects. A significant interaction of two variables indicates that these two variables should be cross-classified and used in the

Table 2. Logistic Regression Results for Method 2 (Bolded Variables were Forced in the Models)¹

Independent Variables	Wald Test	Odd Ratio	Global Test	Deviance Test
1. race, age/sex, tenure	-	-	725.68(11)*	
2. HH Comp 1	87.41(1)**	1.801	820.16(12)*	94.48(1)**
3. HH Comp 2	174.79(1)*	2.240	916.55(12)*	190.87(1)*
4. HH marital status	5.63(1) [0.22](1)	1.205 [1.038]	825.90(13)* [916.769](13)*	5.74(1) [0.22](1)
5. relationship	2.94(1) [5.54](1)	0.911 [0.88]	823.10(13)* [922.09](13)*	2.94(1) [5.54](1)
6. urban	98.90(1)** [102.42](1)**	2.207 [2.244]	905.66(13)* [1004.97](13)	85.50(1)** [88.43](1)**
7. % nonowners	1.65(1) [2.34](1)	1.070 [1.084]	907.31(14)* [1007.32](14)*	1.65(1) [2.35](1)
8. mail response rates (MRR)	30.77(1) [28.37](1)	0.766 [0.774]	935.79(14)* [1032.78](14)*	30.13(1) [27.81](1)
9. % minority	10.96(1) [10.12](1)	0.838 [0.844]	916.64(14)* [1015.11](14)*	10.98(1) [10.32](1)
9. household size (HHS)	25.95(1) [21.92](1)	0.644 [0.668]	929.64(14)* [1025.35](14)*	23.98(1) [20.38](1)

¹Test statistics in the square brackets represent the test statistics obtained from models with HH comp 2.

same dimension of a raking matrix. For example, if 6 age/sex*tenure interactions are jointly significant by deviance test, then age/sex and tenure would lead to form one dimension for raking by cross-classifying age/sex and tenure. Results from interaction testing also help develop collapsing rules for combining the variables for poststratification. Here, we regressed the dependent variable on age/sex, race/origin, tenure, urban, HH comp 2, and two-factor interaction terms of these main effects.

Table 3 presents the results of interaction terms with their main effects and without the main effects. Column

3 shows that interactions are significant when main effects are excluded from the models. On the other hand, column 2 shows that all interaction terms are insignificant when their main effects are included in the models. This is an indication of high multicollinearity between the interaction terms and their main effects. This may also result from using DE of 20.2 to adjust for critical values for significant tests. However, this should not be a concern for developing poststratification scheme for 2000.

Table 3. Logistic Regression Results for Significant Main Effects and Their Interaction Terms

Interaction Terms	Deviance Test (includes main effects)	Deviance Test (excludes main effects)
race X age/sex	47.68 (24)	585.05 (24)
race X tenure	63.02(4)	529.624(4)*
race X HH comp 2	16.51(4)	608.14(4)*
race X urban	4.98(3)	337.83(3)*
age/sex X tenure.	23.39(6)	219.40(6)**
age/sex X HHcomp2	1.45(4)	341.99(4)*
age/sex X urban	13.26(6)	167.98(6)*
tenure X HH comp 2	28.59(1)	579.269(1)*
tenure X urban	1.34(1)	321.65(1)*
HH comp 2 X urban	2.92(1)	527.30(1)*

5. Conclusions

Using the logistic regression analysis, this study identifies that household composition and urbanicity are the only two important variables in addition to race/Hispanic origin, age/sex, and tenure. Urbanicity variable contrasted all urban areas to non-urban areas. Household composition is a person level variable. A person in a housing unit is easy to enumerate if a single person resides in the unit and the person is 50 or older or the unit is occupied by a married couple of 30 years of age or over with 1-5 children of their own under 18. All other persons in the unit are considered hard to enumerate.

Due to different model inference tests and variable selection methods used in this paper, the results are a little different from those of Haines and Hill (1998). Haines and Hill find that relationship and MRR are also important variables. It may be mentioned here that the results may change if one does not force the race/Hispanic origin, age/sex, and tenure variables in the model or different DEs are being used.

References

- Alho, Juha M. Mulry, Mary H. Wurdeman, Kent and Kim, Jay (1993), "Estimating Heterogeneity in the Probabilities of Enumeration for Dual-System Estimation," *Journal of the American Statistical Association*, 88, 1130-1136.
- Haines, Dawn E and Hill, Joan M (1998), "A Method for Evaluating Alternative Raking Control Variables," *American Statistical Association Meetings*.
- Hogan, Howard (1993), "The 1990 Post-Enumeration Survey: Operations and Results," *Journal of the American Statistical Association*, 88, No. 423, 1047-1060.
- Mulry, Mary H., Davis, Mary C., and Hill, Joan M. (1997), "A Study in Heterogeneity of Census Coverage Error for Small Areas," *American Statistical Association Proceedings of the Survey Research Methods Section*.
- Robinson, Gregory J., Ahmed, Bashir, Gupta, Prithwis D. Woodrow, Karen A. (1993), "Estimation of Population Coverage in the 1990 United States based on Demographic Analysis," *Journal of the American Statistical Association*, 88, No. 423, 1061-1071
- Wallace, Leslie and Rust, Keith (1996), "A Comparison of Raking and Poststratification Using 1994 NAEP Data," *Leslie Wallace*, West Inc., 584-589.

DEVELOPING AN AUTOMATED INDUSTRY AND OCCUPATION CODING SYSTEM FOR CENSUS 2000¹

Daniel W. Gillman, Martin V. Appel, Census Bureau
Daniel W. Gillman, Census Bureau, SRD Rm.3000-4, Washington, DC, 20233

Key Words Evaluation, Processing

Abstract *The Bureau of the Census will collect information describing the industry and occupation of jobs of individuals through a series of six questions contained in the long form questionnaire for the 2000 Census. Industry and Occupation (I&O) coding is the process of classifying these responses into 250 industry and 500 occupation categories. It is laborious, expensive, and error-prone, especially the clerical part of the operation. The automated I&O coding system developed for the 1990 Census was very successful. It assigned a code about 48% of the time, thus reducing the labor, time, and errors associated with the coding process.*

Because of the success of the 1990 automated I&O coding system, the Census Bureau decided to use an automated I&O coding system for the 2000 Census. However, the 1990 system was not adaptable to the 2000 Census because of a lack of resources and lack of sufficient research in the intervening years to improve the system. Research on other systems and techniques during that time showed much promise for improvement. Approximately 20 systems were evaluated, and five showed significant promise.

A competitive, three phase development process was adopted for building a 2000 automated I&O coding system. The five vendors that research showed had the systems with the best potential were asked to participate. Each phase involves development by the vendors and independent evaluations of the systems by the Census Bureau. At the end of the first phase, two vendors were asked to participate in the second phase. At the end of the second phase, one or both vendors will continue to develop the production system for the 2000 Census. A fourth phase is the implementation of the 2000 Census I&O coding operation.

This paper describes the three phase development process and its advantages and disadvantages, along

with an explanation of why the 1990 system was not appropriate for 2000. The evaluation and selection processes used for the selection of coding systems to continue in the next phase of development will be discussed. The results of the first phase are presented with a short description of each of the systems. Finally, an evaluation of the current state of the development is given.

Introduction For the 2000 Census, the Census Bureau (BOC) will collect information describing the industry and occupation of jobs of individuals through a series of six questions contained on the long-form questionnaire. The process of classifying these responses is called industry and occupation (I&O) coding. The clerical part of the I&O coding operation is laborious, expensive, and error-prone, so automated I&O coding was introduced for the 1990 Census. The automated I&O coding system (AIOCS) developed for the 1990 Census was very successful. The productivity of AIOCS during the 1990 Census processing was 58% for industry and 38% for occupation at estimated error rates of 6% and 11%, respectively. This capability reduced the labor, time, and errors associated with the coding process.

Because of the success of an integrated coding system, AIOCS and a computer-assisted clerical coder, the BOC decided to use this approach again for the 2000 Census. However, the decision was made not to resurrect AIOCS because of a lack of resources and insufficient research in the intervening years. Evaluations of other systems and techniques during that time showed much promise for improvement. Approximately 20 systems were evaluated, and five showed significant promise.

A competitive, three phase development process was adopted for building a 2000 automated I&O coding system. The phases require development and evaluations by the vendors, independent evaluations by the BOC, development of new test files, and designing the 2000 I&O coding structure and indexes. The five vendors who had the best systems were asked to participate in the first

¹ This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau Publications. This report is released to inform interested parties of research and to encourage discussion.

phase. At the end of the first phase, two vendors were asked to participate in the second phase. At the end of the second phase, one or both vendors will continue to develop the production system for the 2000 Census. A fourth phase is the implementation of the 2000 Census I&O coding operation.

This paper describes the reasons the BOC chose to develop the 2000 Census automated I&O coding system using outside vendors in a competitive, multi-phase development. Included is an explanation of why AIOCS was not appropriate for use in the 2000 Census. The three development phases are described briefly, the criteria for selecting the two systems to participate in Phase 2 of the development are given, and the evaluation technique used to measure the effectiveness of automated coding systems is described. The results of the evaluations of the five systems in Phase 1 are presented, a brief description of the systems is given, and reasons two of the systems did best are offered. Finally, a brief summary of the current state of the development is provided.

Background AIOCS was developed by the BOC (Appel and Hellerman, 1983; Appel and Scopp, 1987) for the 1990 Decennial Census and was declared a success (Scopp and Tornell, 1991). Many factors were involved, but the most important were careful planning and adequate time and funding for the project. This need was recognized early in the 1990 development process.

Unfortunately, it was not possible to use AIOCS for the 2000 Census. For each decennial census, a new I&O coding structure is developed to conform with changes to labor force characteristics and changes in the standard industrial and occupational classification structures. The 2000 industry codes will be based on the new North American Industrial Classification System (NAICS) and the 2000 occupation codes on the vastly updated Standard Occupational Classification (SOC). Much experience, time, and expense are necessary to translate the software and databases in AIOCS to the 2000 Census I&O coding structures and to update the software and databases to conform with new labor force characteristics. The BOC did not have the time and resources available as they had during the 1990 Census development cycle.

Following the 1990 Census, the BOC conducted research into alternative methods of automated I&O coding. First, improving the capability of AIOCS was attempted. This was mildly successful (Gillman and Appel, 1993). Then, an experiment was conducted with Thinking Machines Corp., a previous producer of high-end parallel computers. This was very successful, showing marked

increases in productivity for both industry and occupation coding (Creedy *et al*, 1992; Speizer and Buckley, 1998). However, the company wrote the software in a proprietary language which could not be maintained by the BOC, and the results were never independently verified. The coding technique employed memory-based reasoning (Creedy *et al*, 1992) and was a sharp departure from the artificial intelligence methods (Appel and Hellerman, 1983; Appel and Scopp, 1987) employed in AIOCS.

After this, the BOC began looking at alternative coding techniques used by different organizations around the world. About 20 different systems were considered and evaluated using public-use versions of the evaluation databases that were created for the 1990 development. Of these, the BOC identified five systems which seemed the most promising for developing an automated I&O coder for the 2000 Census. Each of these five systems represented a different technique for performing the coding task, and there were reasons to believe each one could succeed.

Description of Systems The following contains the name of the developer, the country from which the developer comes, the name of the system, and a short description of each of the five systems evaluated in Phase 1.

Inference Group (Australia) - Precision Data. This is a fully automated batch and computer assisted coding system. It is thesaurus based, using natural language text analysis. The system has many parameters which make it easy to change the acceptance criteria for coding.

INSEE (France) - SICORE has been under development since 1992 (Riviere, 1994). It is based on the old QUID system (Lorigny, 1988), developed in the early to mid 1980s. QUID uses information theory along with a tree structure of bigrams (2 letter sequences) of the words in a test file to associate codes with text. Using this tree, information is revealed in a step-by-step fashion; the codes can be determined with partial information. INSEE claims this approach is superior to key word searches, distance measures, and natural language techniques. They have done extensive research to determine which bigrams are most effective in identifying words or codes with their tree. They have done this in both English and French.

National Occupational Safety and Health (NIOSH) (U.S.) - The standard Occupation and Industry Coder (SOIC). This system is composed of three sections: a

rules based artificial intelligence system, an exact match section, and a section based on probabilistic measures to assign a code. The system was originally developed to code death certificates.

Statistics Canada (Canada) - ACTR (Automated Coding by Text Recognition). This is a generalized automated coding system. It can be tailored to fit a variety of automated coding tasks. The software uses a dictionary of phrases, variable input text parsing strategies, and spelling correctors. Its main limitations are that it currently codes single field entries and does best with very short (one or two words) responses.

Wise Enterprises (U.S.) - Dataware Engineering. This system uses a proprietary database to encrypt and compress data. The system uses nearest neighbor searches to rapidly retrieve data. The system can classify to any coding structure for which there is suitable data to train it. The "engine" which drives the coder is commercially available, but the coder is not.

Multi-Phase Approach The BOC developed a four-phase approach for developing and implementing the I&O coding autocoder system for the 2000 Census. The phases are:

Phase 1: (11/15/97 to 05/31/98) - Develop and test five prototypes using 1990 coding structures and evaluation files.

Phase 2: (08/01/98 to 01/31/00) - Develop and test two production-ready systems. Modify autocoders to classify responses to year 2000 I&O codes and to meet decennial census production input, output, and quality assurance (QA) requirements. Perform autocoder tests and improvements, including testing interactive software between the automated and computer-assisted clerical coding systems.

Phase 3: (02/2000 to 04/2000) - Select production system. Integrate autocoder to meet decennial census production input, output, and QA requirements. Perform final production testing and improvements including updating of new industry and occupation phrases, synonyms, abbreviations and improving coder performance from initial decennial responses.

Phase 4: (05/2000 to 03/2001). 2000 Census I&O coding operation.

Selection Criteria The criteria for selecting the vendors to participate in subsequent phases are mainly the

industry and occupation production rates (see Evaluation section below) determined at the target match rates (the minimum acceptable match rate for each code category) of .90 and .87, respectively, using the 1990 Validation File². The better the rates, the better the system. However, if these rates don't differentiate the systems adequately, then speed of the system, price for development, ease of integration and use, and robustness in a network environment are used to help make a decision.

Phase 1 is now complete, the two vendors were selected to participate in Phase 2, and Phase 2 is under way and on schedule. The selection of the two vendors to participate in Phase 2 was easy. The production rate analysis (see Results section below) showed that two systems clearly outperformed the others. Since higher production rates mean less work for the clerks, improved accuracy, and less money spent for the I&O coding operation, the BOC felt the vendors with the two most productive systems had the best chance to build the best systems over time.

Evaluation Each system was evaluated using the same technique employed for AIOCS in the 1990 Census. BOC analyzed the results of coding the 1990 Validation File. The 1990 Validation File was given to the system developers without the "truth" codes. The developers coded it with their systems, and the results were sent back to the BOC for analysis.

The analytical technique, called the **cutoff score method** (Chen, Creecy, and Appel, 1993), estimates productivity by controlling errors for each code category. If the estimated error rate is below a certain level for each code category, then it is below that level over all. If the error rate is measured globally only, then no one is sure if the errors are spread across the code categories or clustered in a few. Provided the overall error rate stays the same, errors clustered in a few code categories render those categories much less useful.

When a system assigns a code to a case, it does so on the basis of a score, which is an estimate of the probability the assigned code is correct. An **acceptable code** is one where the score is above the cutoff score for that code category. The **cutoff scores** are estimated by using a "truth" set such as the 1990 Validation File.

The productivity for each system (calculated for industry and occupation independently) was estimated using six

² 1990 Validation File contains 150 triply coded cases per code category from the 1990 Post Enumeration Survey.

target match rates: 0.86, 0.87, 0.88, 0.90, 0.92, and 0.94. The productivity is the fraction of cases receiving acceptable codes divided by the total number of cases. The target match rate is the minimum acceptable match rate for each code category. The match rate is the fraction of cases with acceptable and correct codes divided by the number of cases with acceptable codes. The match rate is important in comparing systems where the productivity is the same for a given target match rate.

Some cases in the 1990 Validation File don't have "truth" assigned for both industry and occupation. This is an artifact of the way the sample of cases in the file was generated. As a result, estimates of productivity using the entire file are probably too low. On the other hand, if the non-coded case are removed from the analysis, then the estimates of productivity are probably too high. This is due to the under-representation of cases that can't be coded. Therefore, both sets of estimates are presented to provide upper and lower bounds for productivity.

So, the analysis contains two sets of results. The first analysis contains the estimates of productivity and match rates for industry and occupation for the six target match rates for each of the five vendors based on the full 1990 Validation File. The other modified analysis contains the same estimates based on the 1990 Validation File with the non-coded cases removed, i.e. the non-coded industry

cases were removed for the industry analyses and the non-coded occupation cases were removed for the occupation analyses.

Results The 1990 Census autocoder, AIOCS, performed at 58% productivity based on a target match rate of 90% for industry and 38% productivity based on a target match rate of 87% for occupation. The Phase 1 evaluation results of the full 1990 Validation File show that none of the five systems perform at these benchmark levels for both industry and occupation coding. However, the systems from NIOSH and Wise Enterprises exceeded the occupation benchmark using the modified analysis.

The four tables (A-D) below show the results of coding the 1990 Validation file. The first two tables (A,B) contain the analysis of the full 1990 Validation File (69207 cases). The last tables (C,D) contain the analysis for the modified 1990 Validation File, i.e. the file containing just the cases with "truth" codes assigned (64617 cases for industry and 62218 cases for occupation). The columns showing the results for NIOSH and Wise Enterprises are highlighted. Additionally, the target match rate values of 0.90 and 0.87 are highlighted for industry and occupation, respectively. Note: TMR stands for Target Match Rate.

Table A: Industry Production Rates

TMR	Inference	INSEE	NIOSH	Stat Can	Wise
.86	.239	.340	.494	.183	.465
.87	.231	.322	.477	.164	.441
.88	.208	.301	.457	.158	.425
.90	.176	.266	.418	.146	.380
.92	.154	.233	.339	.091	.339
.94	.127	.203	.265	.054	.260

Table A shows that the industry production by NIOSH and Wise Enterprises were first and second for all target match rates as a result of coding the full 1990 Validation

File. However, the production rates for the 0.90 target were below the 1990 AIOCS benchmark.

Table B: Occupation Production Rates:

TMR	Inference	INSEE	NIOSH	Stat Can	Wise
.86	.123	.288	.340	.153	.341
.87	.116	.277	.327	.141	.328
.88	.106	.266	.312	.130	.317
.90	.095	.235	.281	.115	.284
.92	.074	.204	.255	.092	.244
.94	.064	.150	.218	.080	.166

Table B shows that the occupation production by NIOSH and Wise Enterprises were first and second for all target match rates as a result of coding the full 1990 Validation

File. However, the production rates for the 0.87 target were below the 1990 AIOCS benchmark.

Table C: Industry Production Rates:

TMR	Inference	INSEE	NIOSH	Stat Can	Wise
.86	.285	.394	.575	.251	.542
.87	.275	.388	.554	.209	.519
.88	.257	.372	.535	.203	.504
.90	.202	.304	.486	.171	.446
.92	.172	.276	.433	.139	.393
.94	.148	.234	.350	.082	.355

Table C shows that the industry production by NIOSH and Wise Enterprises were first and second for all target match rates as a result of coding the modified

1990 Validation File. However, the production rates for the 0.90 target were below the 1990 AIOCS benchmark.

Table D: Occupation Production Rates:

TMR	Inference	INSEE	NIOSH	Stat Can	Wise
.86	.162	.354	.425	.208	.408
.87	.150	.348	.410	.200	.396
.88	.149	.333	.390	.176	.383
.90	.124	.305	.364	.148	.361
.92	.104	.269	.324	.132	.327
.94	.080	.211	.286	.108	.264

Table D shows that the occupation production by NIOSH and Wise Enterprises were first and second for all target match rates as a result of coding the modified 1990 Validation File. Note that the production rates for the 0.87 target were above the 1990 AIOCS benchmark. These values are further highlighted in the table.

analyzed in two ways, resulting in two sets of rates for each system. (See Evaluation and Results sections for details.) Based on these results, the systems from NIOSH and Wise were approved for selection for Phase 2.

AIOCS was able to code cases at a speed of 10 cases per cpu second. All five systems exceeded that speed. Because computer hardware processing speed has improved so much in the past ten years, this result is not a surprise.

Other criteria were looked at, but the fact that NIOSH and Wise were clearly better than the other three systems made the choice simple. Costs, especially for on-site development and services, were considered, but all vendors submitted similar cost estimates for Phase 2. Speed of the systems was analyzed, but all them substantially exceeded the AIOCS benchmark of 10 cases per cpu second. Finally, ease of integration into the planned Census processing system was considered. NIOSH and Wise both provided systems that were easily integratable.

All the vendors indicated in their final reports that they can improve productivity given more time. The six months duration of Phase 1 was very short. There is every expectation that the vendors can improve their systems substantially. All have identified specific problems they can solve to make these improvements.

The BOC planned to use the additional criteria if the two best systems could not be determined by the production rates alone. However, as the tables show, this was not the case.

Recommendation The results of the analyses of Phase 1 results show that NIOSH and Wise Enterprises appear to have the best systems for I&O coding, based on measuring production rates tuned to certain target match rates. The evaluation test file was

The allotted time for Phase 1 presented numerous difficulties for the vendors. Six months was a short time to develop prototypes that would adequately demonstrate

the capabilities of the systems, but the budget and time constraints placed on the development process made it impossible to expand the schedule. Importantly, each vendor had the same amount of time, and each devoted the resources they thought were necessary to accomplish the task from the beginning. If this process is adopted for future automated coding operations, more time will be needed for each of the phases. AIOCS was under development from the early 1980's through the fall of 1990.

NIOSH and Wise both had previous experience with the 1990 I&O coding structures. NIOSH is building a general purpose automated I&O coder for death certificate coding. This system will be used by the states and the federal government. Wise performed some work in the mid-90's building simple coding prototypes for the BOC. However, Inference Group had the same opportunity with the same data that was given to Wise, but they did not submit I&O coding results for analysis. Statistics Canada and INSEE, national statistical agencies for Canada and France, have had vast experience with automated coding, and each has built many successful coding applications with their software systems.

Conclusion Given the limited time and resources available to automated I&O coding development since the 1990 Census, the BOC decided to hire vendors to build the system for the 2000 Census. Research since the 1990 Census showed there are several competing algorithms that could be successfully applied to the I&O coding problem. So, a three phase development plan was implemented.

Five vendors were asked to participate in the first phase of development. Each vendor uses a different algorithm for coding. Phase 1 is now complete, and two vendors were selected to participate in Phase 2. Phase 2 is under way and on schedule. Preliminary results of Phase 2 indicate the vendors are improving their coding rates monthly. Progress is slow, but the BOC anticipates that one or both of the vendors will exceed the production rates AIOCS achieved in the 1990 Census.

The success of this approach points to a new way

statistical offices can improve specialized statistical computing software such as for automated coding. Expert design and programming skills are hard to retain in the current job market. It is probably cheaper and more effective to build specialized systems "outside" the agency.

References

- Appel, M. V. and Hellerman, E. (1983). "Census Bureau Experiments with Automated Industry and Occupation Coding", *Proceedings of the American Statistical Association*, 32-40.
- Appel, M. V. and Scopp, T. (1987). "Automated Industry and Occupation Coding", presented at *Development of Statistical Expert Systems (DOSES)*, December 1987, Luxembourg.
- Chen, B., Creecy, R. H., and Appel, M. (1993). "On Error Control of Automated Industry and Occupation Coding", *Journal of Official Statistics*, Vol. 9, No. 4, 1993, pp. 729-745.
- Creecy, R. H., Masand, B. M., Smith, S. J., and Waltz, D. L. (1992). "Trading MIPS and Memory for Knowledge Engineering", *Communications of the ACM*, Vol. 35, No.8, 48-64.
- Gillman, D. W., Appel, M. V., (1993). "Analysis of the Census Bureau's Automated Industry and Occupation Coding System Algorithm", *Proceedings of the American Statistical Association*, 1993.
- Lorigny, J. (1988). "QUID, A General Automatic Coding Method", *Survey Methodology*, Vol. 14, No. 2, pp. 289-298.
- Riviere, P. (1994). "The SICORE Automatic Coding System", Working Paper, ISIS 94 Seminar, Bratislava, Slovakia, May 1994.
- Scopp, T. S., Tornell, S. W. (1991). "The 1990 Census Experience with Industry and Occupation Coding," presented at the Southern Demographic Association Annual Meeting, Jacksonville, FL, October 11, 1991.
- Speizer, H. and Buckley, P. (1998). "Automated Coding of Survey Data" in Couper, M. *et al* (eds.) *Computer Assisted Survey Information Collection*, Wiley Series in Probability and Statistics, 1998, pp. 223-243.

SERVICE BASED ENUMERATION ESTIMATION

¹Felipe Kohn and Richard Griffin

Felipe Kohn, Bureau of the Census, Washington DC 20233

Key Word: Multiplicity Estimator

I. Introduction

The Census Bureau established the Service Based Enumeration (SBE) program as the statistical program designed to include persons without usual residence that use service facilities (shelter, soup kitchen or mobile food vans). Those persons are not covered by regular Census Bureau procedures for households or persons in group quarters.

The proposed methodology for the SBE estimation for the 2000 Census is the Multiplicity estimator that is based on the number of times the respondent uses the service facilities.

In this paper we present several multiplicity estimators based on the usage question for service facilities.

II. Estimators - All estimators assume the unduplication of questionnaires has been completed.

Shelter Only Estimator

The enumeration is done only at shelters. The respondents are asked "Including today, how many days in the past seven days have you stayed in a shelter?" Thus, the estimator is obtained as follows:

$$\hat{X}_{sh} = \sum_{i=1}^n 7 / A_i, \text{ where}$$

A_i = shelter usage response for the

i - th respondent and n is the number of respondents

B. Soup Kitchen Estimator

The enumeration is done only at soup kitchens. The respondents are asked "Including today how many days in the past seven days did you eat in a soup kitchen?" Thus, the estimator is obtained as follows:

$$\hat{X}_{sk} = \sum_{i=1}^m 7 / B_i \text{ where,}$$

B_i is the number of days the i -th respondent ate a meal in a soup kitchen and m is the number of respondents.

C. Combined Estimator

The enumeration is done in both service facilities. The estimator is based on the average of the shelter estimator and soup kitchen estimator.

Thus, this estimator is obtained as follows:

$$\hat{X}_{comb} = (\hat{X}_{sh} + \hat{X}_{sk}) / 2$$

D. Optimal Estimator

This estimator is the optimal combination (minimum variance) of the shelter and soup kitchen estimators. It is obtained by the following expression:

$$\hat{X} = \hat{W}_{opt} \hat{X}_{sh} + (1 - \hat{W}_{opt}) \hat{X}_{sk} \text{ where,}$$

$$\hat{W}_{opt} = \frac{Var(\hat{X}_{sk})}{Var(\hat{X}_{sk}) + Var(\hat{X}_{sh})}$$

E. 1995 Type Estimator

The 1995 Census Test used a SBE questionnaire that asked the following question: "How many days during the past week did you use a shelter, soup kitchen or mobile food van?" These questions were asked of all persons at the shelters on the selected day and all persons enumerated in soup kitchens or mobile food vans on the next day. The multiplicity estimator is as follows:

¹Felipe Kohn and Richard Griffin are mathematical statisticians in the Decennial Statistical Studies Division of the U.S. Census Bureau. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

$$\hat{Y}_{1995} = \sum_{i=1}^n 7 / Z_i \text{ where, } n \text{ is the total number of}$$

persons enumerated in a shelter on the selected day and/or a soup kitchen or mobile food van on the next day and Z_i is the number of days a week person i uses a shelter, soup kitchen, or mobile food van.

F. Census Dress Rehearsal Estimator

All persons at a shelter on the selected day are asked "How many days a week including today did you use a shelter?". All persons using a soup kitchen or mobile food van on the next day are asked the same question about soup kitchen or mobile food van usage. They are asked "How many days during the past week did you receive a meal from a soup kitchen or mobile food van?". A person enumerated on both occasions is counted once based on his/her shelter usage. The multiplicity estimator is as follows:

$$\hat{X}_{DR} = \sum_{i=1}^n 7 / A_i + \sum_{i=1}^m 7 / B_i$$

where n is the number of persons enumerated at a shelter on the selected day, m is the number of persons enumerated in a soup kitchen or mobile food van the next day who did not use a shelter during the past week, A_i is the number of days in the past seven days person i uses a shelter and B_i is the number of days in the past week (for those who did not use a shelter) person i received a meal from a soup kitchen or mobile food van.

III. Hypothetical Populations

The persons without usual residence are very transient, by definition. Weather conditions, disposable income, number of service facilities in the area among other factors make it difficult to include some of them in any type of enumeration. In order to take this factor into consideration two pseudo-populations are considered for this study:

For the first four estimators we assume a population of persons without usual residence of 2,500 persons with at least ten percent of them not using services (and thus never enumerated). For the last two estimators we consider a pseudo-population of 3,000 persons with at least 250 never found in a shelter or a soup kitchen and an additional 500 who are never in a shelter but do eat meals in soup kitchens.

IV. Capture Probabilities

Persons without usual residence use service facilities (shelters and soup kitchens) depending on factors such as weather (in rough weather the population in the shelters increases while in mild weather the population in the shelters decreases), income (for example a veteran of the armed forces will not eat in a soup kitchen while he still has money from the VA and will eat in the soup kitchen when his money runs out). Thus, the enumeration of persons without usual residence at a given time varies greatly by the type of service facility (shelter or soup kitchen). To take this factor in consideration, we consider three sets of capture probabilities. For each set the pseudo-populations (3,000 persons and 2,500 persons) are randomly generated using these probabilities (except for persons designated to never use a type of service) assuming independence over days and persons; thus, for each person in the pseudo-population we generate a random number (between 0 and 1) for each day in a week for the shelter usage and a random number for each day in a week for the soup kitchen usage. If the generated random number on a given day is less than the set capture probability the person is designated as using the service facility that day. Three sets of capture probabilities are considered (a new set of random numbers is generated for each set of capture probabilities):

- A. 50% probability of using a shelter and 50% probability of using a soup kitchen.
- B. 50% probability of using a shelter and 25% probability of using a soup kitchen.
- C. 25% probability of using a shelter and 50% probability of using a soup kitchen.

V. Statistical Criteria

- A. The selected statistic to compare the estimates is the relative root mean square error defined by the following expression:

$$RRMSE = \frac{\sqrt{MSE(\hat{X})}}{X}$$

with X equal to the true population size and

$$MSE(\hat{X}) = Bias(\hat{X})^2 + Var(\hat{X})$$

B. Variance Estimation

The methodology for determining the variance for all the multiplicity estimates presented in this paper was derived by Pat Cantwell of the Census Bureau and is as follows: Divide the target population into eight groups: G_0, \dots, G_7 , of sizes N_0, \dots, N_7 , where the N_i 's are the number of persons who use a shelter (soup kitchen) i days in a week. Assume that the days a person uses a shelter (soup kitchen) are selected randomly given the number of days a week he/she uses a shelter.

N is the total target population. On the selected enumeration day n_i out of N_i use a shelter for $i=1, \dots, 7$ with each n_i distributed as a binomial random variable with parameters N_i and $i/7$.

The multiplicity estimator can be written as

$$\hat{N} = \sum_{i=1}^7 \frac{7}{i} n_i$$

$E(\hat{N} | \{N_0, \dots, N_7\}) = N - N_0$ under our assumptions and the variance is as follows:

$$\begin{aligned} \text{Var}(\hat{N}) &= \sum_{i=1}^7 (7/i)^2 (N_i)(i/7)(1-i/7) \\ &= \sum_{i=1}^7 N_i \left(\frac{7-i}{7} \right) \end{aligned}$$

This binomial model allows for variation in which persons are at a shelter on the enumeration day and fixes the partition of the population into N_1, N_2, \dots, N_7 . For this model the variance can easily be calculated allowing the N_i values to vary. A more detailed description of this less restricted variance can be read in the paper presented at this session by Roger Shores of the Census Bureau.

VI. Results

A. Tables 1-3 below shows the results of our simulations for the estimators based on a pseudo-population of 2,500.

Table 1: Estimators with Capture Probabilities of $P=.5$ for Shelters and $Q=.5$ for Soup Kitchens

Estimator	Estimate	Variance	C.V.	RRMSE
Shelter Only	2235	3008.7	2.45%	10.80%
Soup Kitchen	2237	3194.9	2.52%	10.76%
Combined	2236	1550.9	1.76%	10.67%
Optimal	2236.28	1549.51	1.76%	10.64%

Table 2: Estimators with Capture Probabilities of $P=.5$ for Shelters and $Q=.25$ for Soup Kitchens

Estimator	Estimate	Variance	C.V.	RRMSE
Shelter Only	2223	3109	2.50%	11.30%
Soup Kitchen	1935	6360	4.12%	22.82%
Combined	2079	2367.42	2.34%	16.95%
Optimal	2205.81	2088.51	2.07%	11.9%

Table 3: Estimators with Capture Probabilities of $P=.25$ for Shelters and $Q=.5$ for Soup Kitchens

Estimator	Estimate	Variance	C.V.	RRMSE
Shelter Only	2250	980.38	1.39%	10.09%
Soup Kitchen	2234	3127.21	2.5%	10.87%
Combined	2242	1026.9	1.42%	10.39%
Optimal	2246.28	746.38	1.21%	10.21%

B. Table 4 below, shows the statistics of the simulation for the second pseudo-population.

Table 4: Statistics for the 95 Test and 2000 Dress Rehearsal Estimators

Estim.	Prob	Est	Var	C.V.	RRMSE
2000 Dress Reher.	P=.5 Q=.5	2747	3949.75	2.28%	8.70%
	P=.5 Q=.25	2664	4812	2.60%	11.43%
	P=.25 Q=.5	2747	1726.58	3.06%	8.88%
95 Census Test	P=.5 Q=.5	2747	1685.32	1.49%	8.54%
	P=.5 Q=.25	2664	3246.55	2.13%	11.51%
	P=.25 Q=.5	2747	1126.58	1.22%	8.61%

C. In table 5 we present results of the SBE estimation from the 2000 Census Dress Rehearsal in Sacramento and in Columbia, S.C. It is important to note that in Columbia we did not do any estimation but since all the questionnaires and field procedures were identical to their counterparts in Sacramento, we applied the estimation procedures for research purposes only.

Table 5: Results from the 2000 Census Dress Rehearsal

Site	Persons enumerated		Estimate
Sacramento	Shelter	584	2443
	Soup Kitchen	500	
Columbia	Shelter	254	833
	Soup Kitchen	187	

VII. Analysis

In all the estimators the relative mean square error is relatively large (in comparison with the coefficient of variation). The reason for this is the considerable number of persons without usual residence to whom we gave a zero probability of using service facilities.

The shelter only estimator or the soup kitchen only estimator worked fine provided that most of the persons without usual residence use that type of facility. However, based on the results of the 2000 Census Dress Rehearsal, many persons may use soup kitchens but not shelters or shelters but not soup kitchens. A large number of the persons enumerated in soup kitchens answered that they did not use a shelter in the past week (for example 80% of those enumerated in soup kitchens in Sacramento answered that they did not use shelters in the past week).

Comparing the numbers between the 95 test estimator and the 2000 Census Dress rehearsal estimator, the Cvs and the RRMSEs from the 95 Test estimator produce slightly smaller values than their counterpart of the Dress Rehearsal estimator. The main difference between these two estimators is likely the response bias (not considered in this paper) in the 95 test estimator. For example, if a respondent has eaten three days in a soup kitchen and slept in a shelter the same three days, the correct answer to the 95 test question is three days but he or she may answer six. Considering all these estimators from a total error point of view, we feel that the 2000 Dress Rehearsal estimator is best and it is proposed for Census 2000.

MISSING DATA IN THE U.S. CENSUS 2000 DRESS REHEARSAL - AN OVERVIEW

Steven P. Hefter, Lisa D. Fairchild, Philip M. Gbur, U.S. Census Bureau
Steven P. Hefter, U.S. Census Bureau, Washington, DC 20233

Key Words: Data Quality, Item Nonresponse, Imputation, Allocation

I. Introduction

The U.S. Census Bureau conducted the Census 2000 Dress Rehearsal (DR) in 1998 in Sacramento, CA; Menominee, WI; and Columbia, SC and surrounding counties. In the Columbia site we used components of a traditional census methodology which included a post-enumeration survey (PES). The DR PES was similar in design to the Integrated Coverage Measurement (ICM) Survey used in the Sacramento and Menominee sites where a sampling census methodology was employed. As with any census or survey, missing data was encountered throughout the process. This paper gives a brief overview of census operations including the initial phase, the ICM/PES, and the estimation methodology and the levels of missing data encountered.

II. Initial Phase

A. Operations

The Initial Phase operations included creating a list of the addresses in the three sites; an enumeration of households, group quarters and persons without a usual residence; followup for nonresponse; and the data capture. In each site, people in all residential addresses were given an opportunity to mail back a questionnaire. Those who did not mail back a questionnaire by the cutoff date were included in nonresponse followup (NRFU). Traditionally, enumerators are sent out to all nonresponding households, but sampling for NRFU was done in Sacramento for blocks not selected in the ICM sample. All nonresponding housing units (HUs) in ICM blocks were followed up in the field [1]. This paper discusses data collected by the nonresponse followup enumerators but does not include estimation of nonsampled nonrespondents [2].

B. Missing Data Procedures

Even after completion of NRFU, sample or otherwise, census questionnaires may still have missing data.

A questionnaire may not have enough information to determine whether or not the HU is occupied. These are defined to be unclassified returns. Unclassifieds were included in NRFU estimation and thus are not discussed here. The methodology in [2] includes estimation for unclassifieds.

Classified returns must have at least a HU status (occupied, vacant, or delete from the universe) and, if the HU is occupied, the number of persons in the household. Other questionnaires may be complete except for a few questions, referred to as item nonresponse. In addition, as part of the data processing, the questionnaire responses are subject to edits which may result in a response being changed. The assignment of values to complete the questionnaire is called allocation. The allocation procedures used include the following: 1) determination of an appropriate response based on other reported data (such as calculating age from date of birth); 2) assignment of a value based on characteristics of people with similar values for related characteristics (hot deck imputation); and 3) complete substitution of data for a person or all people in a household from a nearby person or household.

If the amount of data provided for an individual is sufficient, then the person is coded as data defined. For the Dress Rehearsal, a person in a HU had to have at least two of the following characteristics to be considered data defined: name, relationship, sex, date of birth or age, Hispanic origin, or race. One or more persons in a household may not be data defined. If all persons in a HU are not data defined then the whole household is substituted [3].

C. Results

1. Allocation

Allocation percentages are provided in Tables 1-3 by DR site for all people (includes persons in HUs, persons in group quarters, and persons without a usual residence).

Allocation percentages for sex are relatively low and assignment of values was often based upon reported data. The percentages range from 5.0 for South Carolina to 7.1 in Menominee. There is little difference in the percentages across each of the three sites.

The age allocation percentages are quite high, but age could often be determined from reported data

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

(particularly, date of birth). The percentages range from 25.4 for Sacramento to 41.4 for Menominee but, excluding cases where age was calculated from date of birth, the percentages range from 6.7 for South Carolina to 8.7 for Sacramento. Since age is used only in groupings for estimation, any methodology that is reasonable should have minimal impact on estimation. There is little difference in the latter percentages across the three sites.

The Hispanic Origin allocation percentages range from 8.9 for South Carolina to 10.4 percent for Menominee. However, this allocation is not just for Hispanic/NonHispanic but also includes type of Hispanic. These percentages are also similar for each site.

Race allocation percentages for all people range from 4.6 for South Carolina to 10.2 for Sacramento. The race allocation percentages are particularly high for the Sacramento Hispanic poststratum. They run from 25 to 30 percent. However, since the poststratum assignment is dictated by the Hispanic origin item, this would not affect estimation.

2. Data Defined Percentages

Percentages of non-data defined persons are provided in Table 4 by DR site for all people. The total percentages range from 3.4 for South Carolina to 5.4 in Menominee. One or more people in a household may be non-data defined (within household) or all people in a household may be non-data defined (whole household). Of these two types, the within household percentage is higher in Sacramento and Menominee while the percentages of within household and whole household are the same in South Carolina.

Data defined percentages are provided in Tables 5-7 by poststratum race and DR site for all people. The "Non-Data Defined: Imputed" columns include non-data defined persons within households. The "Non-Data Defined: Substituted" columns include persons in whole household substitutions. Comparisons of the data defined percentages with the substituted percentages provide a measure of the results for the whole household substitutions while comparisons of the data defined and total percentages provide an indication of the overall effect of the whole person and whole household imputations or substitutions.

Overall, the largest percentage point differences were seen in South Carolina where 57.3 percent of data defined people are White/Other while only 46.1 percent of persons in whole household substitutions were

White/Other with corresponding balancing changes in the percent Black. However, since only 1.7 percent of all people were allocated via whole household substitutions, the differences between the total and data defined person percentages of White/Other and Black are 0.6 percentage points.

Shifts in the distributions between data defined and whole household substituted people are also seen in Sacramento and Menominee. In Sacramento, there is a decrease in the percentage White/Other and Asian with increases in percent Black and Hispanic for the substituted people. For Menominee, there is a decrease in the percentage American Indian with an increase in percent White/Other and Hispanic.

The impact of the somewhat skewed racial distributions of the substituted persons has minor impact on the site-level racial distributions. However, the impact could be pronounced at smaller geographic levels.

D. Conclusions

In general, the missing data procedures for the initial phase should not have a significant effect on the site-level estimation. Further research is needed to investigate the effect at smaller geographic levels.

III. ICM/PES

A. Sample Design

The Dual System Estimation (DSE) methodology used in the DR to measure and account for coverage error requires a second, independent enumeration of persons. All addresses in blocks selected for inclusion in the ICM/PES were included in an address listing separate from the census address listing. A cluster sample of HUs was taken from groups of blocks formed for the ICM/PES [4]. Large and small block clusters were subsampled at varying rates to balance field work concerns with sample size considerations. A total of 1,085 block clusters containing 34,890 HUs were included in sample for the DR ICM/PES. This sample is called the P-Sample. The results from the initial phase in these same block clusters are used in conjunction with the P-Sample, and is referred to as the enumeration or E-Sample.

B. Data Collection and Processing

Enumerators were sent to the clusters in sample. The HUs listed in these clusters during the independent listing made up the P-Sample. The initial phase HUs in the same clusters made up the E-Sample. The P and E-Sample persons went through a matching process to determine whether the P-Sample persons were residents

on census day and whether they match to the E-Sample. For E-Sample persons it was determined whether they were correctly or erroneously enumerated in the initial phase of the census [5].

C. Missing Data Procedures

1. Overview

The ICM/PES missing data system was independent from initial phase missing data procedures and accounted for noninterviews and imputed missing responses. The ICM/PES missing data system also calculated a residence probability (for persons with an unresolved residence status) and a match probability (for those persons with an unresolved match status) [6].

2. Whole Household Noninterviews

The ICM/PES missing data system accounted for whole household noninterviews (NIs) in the P-Sample with a NI adjustment. This NI adjustment proportionally redistributed the P-Sample weights of the noninterviewed HUs to the interviewed HUs within block cluster and Type of Basic Street Address.

3. P-Sample Person Characteristic Imputation

The DSE methodology used in the DR required each person in the P-Sample to have sufficient data to be placed in a poststratum. The variables eligible for imputation for P-sample people were race, Hispanic origin, sex, tenure, and age. In the following discussion, the term "previous" refers to a household processed prior to the one in question. Missing tenure was imputed from the closest previous household having the same TOBA. Missing race was imputed from the race distribution within the household. If everyone in the household had a missing value of race, then the nearest previous household, having similar Hispanic origin was used. Hispanic origin was imputed analogously to race. Missing age was imputed from the distribution of age for persons with a similar relationship to, and age of, the reference person. Missing age for single person households was imputed from the age distribution within all single person households. Missing sex was imputed to be the opposite of the spouse's (with spouse present). In households where there was a reference person and a spouse, and both had missing sex, the reference person's sex was imputed from the sex distribution for persons in households where their spouse was present. The spouse's missing sex was then imputed to be the opposite of the reference person's. For all other persons with missing sex, the sex distribution within similar households was used to impute sex for the reference person [6].

4. E-Sample Enumeration Status

The DSE methodology requires that all E-Sample persons be classified as either correctly or erroneously enumerated. E-Sample people with an unresolved enumeration status were assigned a probability of being correctly enumerated in the census. This probability was computed within DR site and before-follow-up match code group as the simple weighted proportion of correct enumerations (among those persons with a resolved final enumeration status).

D. Estimation Methodology

The person matching results are used in calculating the DSEs for each of the 84 poststrata (6 Race \times 7 Age/Sex \times 2 Tenure) in each DR site. The DSEs are then placed into a two dimensional matrix (Race by Age/Sex \times Tenure). These cell counts are summed to form the marginal constraints. The initial phase estimates for each poststrata were placed in the interior cells of the matrix and the iterative proportional fitting methodology, commonly referred to as raking, was used to force the initial phase estimates to the marginal totals, minimizing the variances and providing results that were consistent across poststrata [7]. The raked DSEs were divided by the initial phase results to yield 84 coverage factors (one for each poststratum) which were used in the subsequent small area estimation [8].

The ICM results were used in the Sacramento and Menominee sites not merely as an evaluation tool, but were incorporated into the final DR estimates. In South Carolina, the official counts did not include the PES results.

E. Results

1. Noninterview Percentages

A measure of quality for any survey is the number of NIs. However, depending upon the purpose or use of the number, there may be multiple definitions for the percent NI, and this is the case with the DR ICM/PES. The percentages for the DR could vary by three criteria: 1) treatment of vacant HUs; 2) treatment of HUs with a preliminary outcome code of "10"; and 3) use of preliminary versus final outcome codes. For purposes of field operations, vacant HUs are included in the denominator of the percent NI. However, for population estimation purposes, vacants are excluded from the denominator. A preliminary outcome code of "10" represents "No census day residents." In Menominee, this code was erroneously applied to many seasonal vacant HUs that should have received a code of "11" ("Vacant on census day"). Thus, it was decided to convert HUs with an outcome code of "10" to "11." To

ensure consistency across sites, and since the same error may have occurred in the other sites, the conversion was applied to all three DR sites. Note that during final outcome code processing, some additional "10"s may be created - these are left as code "10".

Table 8 includes the components used in calculating the noninterview percentages by site and reflects the final treatment of HUs with a preliminary outcome code of "10". In Menominee, the misclassification of seasonally vacant HUs as noninterviews, if left unaddressed, would have substantially misrepresented the quality of the ICM survey in the site. The final estimation NI percent for Menominee would be 30.9 when including the 176 units with preliminary outcome codes of "10". After these units were reclassified as vacant, the final NI percentage drops significantly to a rate of 1.7. The final estimation NI percentages for Sacramento and South Carolina were 5.1 and 5.2 respectively.

2. P-Sample Characteristic Imputation Percentages

Missing data rates for P-Sample persons by DR site are given in Table 9. The selected variables were eligible for allocation. The item nonresponse percentages range from a high of 2.2 percent for age and race in Sacramento to a low of 0.0 percent for Hispanic origin in Menominee. As is to be expected with a coverage survey such as the ICM, all levels of selected item nonresponse were generally very low. Over all three sites, sex had the lowest allocation percentages ranging from 0.1 in Menominee to 0.4 in Sacramento and South Carolina. The high response rate can most likely be attributed to the ease with which the enumerator can determine sex during the interview.

Race and age item nonresponse percentages are relatively consistent across all three sites. Roughly 2.0 percent of all P-Sample persons were allocated these variables. Tenure allocation was also fairly uniform and exceedingly rare in the DR and ranged from 0.2 percent to 0.6 percent. Hispanic origin item nonresponse was generally very low. In Menominee 0.0 percent of the P-Sample people were allocated Hispanic origin. Roughly 1.0 percent of the P-Sample people in Sacramento and South Carolina were allocated Hispanic origin data item.

3. E-Sample Enumeration Status Percentages

Table 10 provides the final E-Sample unweighted percentages for each of the four enumeration status match code groups by site. The four groups are correct enumerations, unresolved erroneous enumerations, erroneous enumerations, and persons with insufficient information for matching. The correct enumeration

percentages across all three sites were fairly consistent ranging from 86.4 in South Carolina to 88.5 percent in Menominee. The Sacramento site had the highest percentage of people with unresolved enumeration status (3.6) and insufficient information (3.7). The percentages of people with unresolved enumeration status and insufficient information ranged from 0.5 in Menominee to 3.6 in Sacramento, and 1.2 in Menominee to 3.7 in Sacramento, respectively. Menominee had a high rate of erroneous enumerations at 9.7 percent, possibly due to a high incidence of geocoding error. The range of percentages for persons with an erroneous enumeration match code is 6.2 to 9.7.

F. Conclusions

As with the initial phase, the handling of missing data did not have a large impact on the ensuing site-level estimation. The amount of missing P-Sample data was small relative to the number of HUs and people interviewed.

In tables 1-10 percentages may not sum due to rounding and the following abbreviations may appear:

NA - Not Applicable

AI - American Indian/Alaska Native

NH/PI - Native Hawaiian/Pacific Islander

Hisp - Hispanic origin

HH - Household

Table 1: Sacramento Allocation Percentages by Characteristic and Allocation Component

Data Item	Base	Percent	Allocation Percent Components		
			Reported Data	Hot Deck / Consistency	Substituted
Sex	354105	6.8	1.4	3.2	2.2
Age	354105	25.4	16.8	6.5	2.2
Hispanic Origin	354105	9.8	0.6	7.0	2.2
Race	354105	10.2	NA	8.0	2.2

Table 2: South Carolina Allocation Percentages by Characteristic and Allocation Component

Data Item	Base	Percent	Allocation Percent Components		
			Reported Data	Hot Deck / Consistency	Substituted
Sex	647896	5.0	1.4	1.9	1.7
Age	647896	28.5	21.8	5.0	1.7
Hispanic Origin	647896	8.9	0.2	6.9	1.7
Race	647896	4.6	NA	2.9	1.7

Table 3: Menominee Allocation Percentages by Characteristic and Allocation Component

Data Item	Base	Percent	Allocation Percent Components		
			Reported Data	Hot Deck / Consistency	Substituted
Sex	4535	7.1	1.3	3.9	1.9
Age	4535	41.4	34.0	5.6	1.9
Hispanic Origin	4535	10.4	0.1	8.4	1.9
Race	4535	6.0	NA	4.1	1.9

Table 4: Non-Data Defined Person Percents by Dress Rehearsal Site and Type

Dress Rehearsal Site	Base	Non-Data Defined Person Percent		
		Total	Within HH	Whole HH
Sacramento	354105	5.1	2.9	2.2
South Carolina	647896	3.4	1.7	1.7
Menominee	4535	5.4	3.6	1.9

Table 5: Sacramento Data Defined Percents by Poststrata Race

Race	Total		Data Defined		Non-Data Defined: Imputed		Non-Data Defined: Substituted	
	#	%	#	%	#	%	#	%
Total	354105	100.0	336088	100.0	10392	100.0	7625	100.0
White/Other	157630	44.5	152279	45.3	2531	24.4	2820	37.0
Black	54953	15.5	51547	15.3	1780	17.1	1626	21.3
AI	10943	3.1	10343	3.1	286	2.8	314	4.1
NH/PI	2482	0.7	2294	0.7	143	1.4	45	0.6
Asian	57115	16.1	53116	15.8	3097	29.8	902	11.8
Hisp	70982	20.0	66509	19.8	2555	24.6	1918	25.1

Table 6: South Carolina Data Defined Percents by Poststrata Race

Race	Total		Data Defined		Non-Data Defined: Imputed		Non-Data Defined: Substituted	
	#	%	#	%	#	%	#	%
Total	647896	100.0	625804	100.0	11246	100.0	10846	100.0
White/Other	367612	56.7	358808	57.3	3805	33.8	4999	46.1
Black	257772	39.8	245389	39.2	6928	61.6	5455	50.3
AI	3520	0.5	3366	0.5	97	0.9	57	0.5
NH/PI	382	0.1	369	0.1	7	0.1	6	0.1
Asian	6209	1.0	5948	1.0	124	1.1	137	1.3
Hisp	12401	1.9	11924	1.9	285	2.5	192	1.8

Table 7: Menominee Data Defined Percents by Poststrata Race

Race	Total		Data Defined		Non-Data Defined: Imputed		Non-Data Defined: Substituted	
	#	%	#	%	#	%	#	%
Total	4535	100.0	4288	100.0	163	100.0	84	100.0
White/Other	581	12.8	561	13.1	6	3.7	14	16.7
Black	4	0.1	4	0.1	0	0.0	0	0.0
AI	3831	84.5	3617	84.4	152	93.3	62	73.8
NH/PI	1	0.0	1	0.0	0	0.0	0	0.0
Asian	2	0.0	2	0.0	0	0.0	0	0.0
Hisp	116	2.6	103	2.4	5	3.1	8	9.5

Table 8: ICM/PES Noninterview Components based on Final Outcome Codes by Site

Component	Sacramento	Menominee	South Carolina
Total Addresses	16419	794	17677
A. Interview	14322	409	14972
B. NI - refusal, no one home, etc.	486	2	495
C. NI - no data defined people	93	0	66
D. No census day residents (10)	186	5	261
E. Vacant	1118	368	1208
F. Not a HU	214	10	675
Final Estimation NI Rate*	5.1	1.7	5.2

* NI Rate Definition: $(B + C + D) / (A + B + C + D)$

Table 9: ICM/PES P-Sample Person Missing Data Percentages for Selected Variables by Site

Variable	Sacramento	Menominee	South Carolina
Base *	36336	1271	35920
Sex	0.4	0.1	0.4
Age	2.2	1.6	2.2
Hispanic Origin	1.3	0.0	1.1
Race	2.2	1.0	1.8
Tenure	0.6	0.2	0.5

* Base excludes people with a residence status code of "Remove."

Table 10: ICM/PES E-Sample Final Unweighted Correct Enumeration, Unresolved Enumeration Status, Erroneous Enumeration, and Insufficient Information Counts and Percents by Site

Status	Sacramento	Menominee	South Carolina
Base	35806	1202	33959
Correct Enumeration Percentage	86.5	88.5	86.4
Unresolved Erroneous Enumeration Status Percentage	3.6	0.5	2.4
Erroneous Enumeration Percentage	6.2	9.7	9.1
Insufficient Information Percentage	3.7	1.2	2.1

REFERENCES

- [1] Singh, R., Cantwell, P. and Kostanich, D. "Census 2000 Dress Rehearsal Methodology and Results," Presented at the American Statistical Association Joint Statistical Meetings, August 10, 1999.
- [2] U.S. Census Bureau, "Specifications for Nonresponse Followup Sampling and Undeliverable-as-Addressed Vacant Sampling for the Census 2000 Dress Rehearsal," internal memorandum for Lynch from Singh, DSSD Census 2000 Dress Rehearsal Memorandum Series #A-35, April 10, 1998.
- [3] U.S. Census Bureau, "1998 Dress Rehearsal Census 2000 One Hundred Percent Imputation Specifications, Version 3," Population Division internal document, September 3, 1998.
- [4] U.S. Census Bureau, "Computer Specifications for the Selection of the ICM Sample for the Census 2000 Dress Rehearsal", internal memorandum for Lynch from Kostanich, DSSD Census 2000 Dress Rehearsal Memorandum Series #A-5, November 14, 1997.
- [5] U.S. Census Bureau, "The Design of the Census 2000 Dress Rehearsal Integrated Coverage Measurement", internal memorandum for Ramos from Childers, DSSD Census 2000 Dress Rehearsal Memorandum Series, Chapter F-DT-2, December 15, 1998.
- [6] Ikeda, M and Kearney, A., "Handling of Missing Data in the Census 2000 Dress Rehearsal Integrated Coverage Measurement Sample", Presented at the American Statistical Association Joint Statistical Meetings, August 10, 1999.
- [7] Schindler, E., "Iterative proportional fitting in the 2000 Census Dress Rehearsal", Presented at the American Statistical Association Joint Statistical Meetings, August 10, 1999.
- [8] U.S. Census Bureau, "Census 2000 Dress Rehearsal Computer Specifications for the Integrated Coverage Measurement Block Level Estimation," internal memorandum for Stoudt from Griffin, DSSD Census 2000 Dress Rehearsal Memorandum Series #A-86.

ERROR PROFILE FOR THE CENSUS 2000 DRESS REHEARSAL

Susanne L. Bean, Katie M. Bench, Mary C. Davis, Joan M. Hill, Elizabeth A. Krejsa, David A. Raglin,
U.S. Census Bureau

Joan M. Hill, U.S. Census Bureau, Room 1002-F FOB 2, Washington, DC 20233

Keywords: Integrated Coverage Measurement/Post Enumeration Survey, Matching Error Study, Evaluation Followup Interview, Data Collection Mode Study, measurement error.¹

1. INTRODUCTION

The error profile examines specific sources of error corresponding to the Census 2000 Dress Rehearsal Integrated Coverage Measurement/Post Enumeration Survey (ICM/PES) that are feasible to measure given the design of the ICM/PES. A sample of ICM/PES block clusters in each site was selected (187 total block clusters across three sites) to assess the magnitude of nonsampling error. This is known as the evaluation cluster sample. The errors with regard to the 'one-number census' in Sacramento, CA and Menominee, WI may occur in the initial dress rehearsal enumeration operation (i.e., initial phase), the ICM enumeration (i.e., final phase), or both. Similarly, the errors measured within the South Carolina site may be found in both the census enumeration and the PES activities. In all three sites, the objective of the error profile is to measure error in the ICM/PES process.

The individual sources of error that are isolated and examined separately in this report are data collection (in both the E-sample and the P-sample²) and instrument error, certain errors in the processing of data (the focus here is errors from the ICM/PES clerical

matching operation), and the effects of alternative data collection modes.

These survey measurement and processing errors are evaluated using the following three tools: Matching Error Study, Evaluation Followup Interview, and the Data Collection Mode Study.

Although production and evaluation operational problems made it impossible to conduct any of these studies as originally intended, the error profile evaluation yielded some interesting results.

2. MATCHING ERROR STUDY

One source of processing error in the ICM/PES is clerical person matching error. People collected in the ICM/PES in a cluster are matched to people found by the initial phase in the same cluster. The first step in this process is a computer match, where obvious matches are made and possible matches are identified. The possible matches and remaining nonmatches are then matched clerically to find the less obvious matches, first by lower-level matchers, and then by matching experts.

The Matching Error Study (MES) attempts to measure the error in the clerical matching process by having expert matchers rematch persons within each block cluster in the evaluation cluster sample. The results from the rematching operation are compared to the production results to find differences in match status.

The discrepancy rates between the production and ICM/PES matching operations were less than one percent in each of the three sites: Sacramento, South Carolina, and Menominee. Presumably they would have been higher if the matching experts had not performed a 100 percent quality assurance during the production matching operation.³ However, the

¹This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

²An E-sample housing unit is a housing unit which is counted in the initial phase at the time the person matching begins and is in an area included in the ICM/PES sample (Childers, 1998). A P-sample housing unit is one that is listed in the ICM/PES listing book and is confirmed to exist in a block cluster in an ICM/PES sample area.

³The original design of the MES was to use matching experts to do a sample quality assurance review of the work done by the clerks and technicians. However, due to last minute production changes which resulted in the

relatively small matching error does suggest that the matching expert coding is highly reliable.

According to the Census 2000 design, after matchers have passed an initial 100 percent quality assurance, matching experts will perform quality assurance on only a sample of cases during production matching. Therefore, the Census 2000 study of matching error is expected to measure the actual magnitude of matching error in the Accuracy and Coverage Evaluation and its subsequent effect on the Census 2000 Dual System Estimation.

3. EVALUATION FOLLOWUP INTERVIEW

The Evaluation Followup Interview (EFU) measures aspects of two types of survey error. The first type is measurement error, the error introduced into the survey process by the interviewer, respondent, and instrument. That error is measured by the Evaluation Person Followup Reinterview, a replication of the ICM/PES Person Followup Interview in a subset of the clusters in the evaluation cluster sample. The production ICM/PES Person Followup Interview is conducted when people from the initial phase and the ICM/PES do not match after the initial clerical person matching operation. The interview collects information to ensure that all correct matches are made and correct residence status are set.

This evaluation attempts to obtain the true match and residence status by giving the clerical matchers a second set of data from the Evaluation Person Followup Reinterview, along with the production Person Followup Interview data, to use when determining the final match and residence status of each person. The comparison of these results with the production data provides an estimate of measurement error in the ICM/PES data.

The second type of error the EFU attempts to measure is production error due to the decision to not conduct a Person Followup Interview for certain people who did not match between the initial phase and the ICM/PES. This study is called the Person Followup Criteria Evaluation. Research in previous Census tests suggested not including these people in the Person Followup Interview, but to code them as residents in the cluster because it was doubtful that useful

information would be gleaned from the Person Followup Interview.

The underlying assumption of the Evaluation Followup analysis is that the EFU process results in residence status and match codes that are closer to truth than results of the ICM/PES process. There were several operational components aimed at meeting this assumption: 1) The field operation was conducted using experienced interviewers, 2) An automated system was used to print person workload information and form questions simultaneously for the PFU Criteria Evaluation, thus eliminating error associated with clerical transcription/preparation of the EFU forms, 3) matching experts were used in lieu of matching clerks to assign match codes and residence status, and 4) information from all previous contacts was used to resolve evaluation cases.

Despite the best efforts of the evaluation staff to design the EFU to produce accurate error estimates, limitations should be considered when interpreting results. Because of processing demands, the EFU interview was not in the field until approximately eight months after Census day. The ability to establish "truth" for the portion of the population that we are most concerned with eight months after Census day is highly questionable due to recall error. Time lag limitations also apply to the ICM/PES production results, since the ICM/PES Person Followup interview directly precedes the Evaluation Followup Interview. As a result, error attributable to the ICM/PES may be understated. An additional limitation of the EFU Interview is the data collection mode, that is, paper and pencil interview as opposed to a Computer Assisted Personal Interview (CAPI). The paper form limits the ability to ask complex question skip patterns, which could be implemented with a CAPI instrument. The decision to use a paper instrument was driven by unavoidable timing and resource restrictions. In addition to these limitations, the EFU could not be conducted as originally intended and certain analysis could not be performed due to operational problems with the evaluation.

3.1 Evaluation Person Followup Reinterview

If we assume the EFU data are closer to the truth than the production data, residence status changes from ICM/PES to EFU indicate that ICM/PES was in error. To evaluate the magnitude of measurement error introduced by the survey process, we produced crosstabulations of the residence status of each person

matching experts' review of every cluster, the focus of the study shifted to an assessment of matching reliability.

as determined after the EFU versus the final ICM/PES residence status. When there was a conflict, information collected in the EFU was used to help determine the source of such conflict and the resolution. Based on these crosstabulations, gross difference rates are estimated.

The gross difference rate (GDR) (Forsman and Schreiner, 1991) is the proportion of people whose ICM/PES and EFU assigned codes differed. The GDR is calculated by dividing the total number of conflicts by the total number of cases. Interpretation of the GDR is subjective.

Regarding match codes, the GDR was 9.7 percent (SE=4.6 percent) in Sacramento and 7.5 percent (SE=3.1 percent) in South Carolina. For the residence status, the GDR was 15.2 percent (SE=5.9 percent) for Sacramento and 9.0 percent (SE=2.8 percent) in South Carolina. For Menominee, the GDR was 11.3 percent (no standard error estimate was produced) for both match and residence status.

These results are in the range of moderate concern, but given the reduced sample of clusters for the Evaluation Person Followup Reinterview due to evaluation operational problems (49 of the 187 original evaluation block clusters), no specific conclusions can be made from these results. The Census 2000 Evaluation Followup Interview design will take these results into consideration.

3.2 Person Followup Criteria Evaluation

The Person Followup (PFU) Criteria Evaluation was conducted using the EFU form which was a modified Person Followup Interview form. It collected information about all people in the evaluation sample clusters who did not match initial phase people but were excluded from the Person Followup Interview. These results were compared to the production results with regard to changes to match and residence status codes.

In addition, the PFU Criteria Evaluation results were used to recompute the Dual System Estimates (DSEs) which are compared to the production DSEs (based on ICM/PES production data from within the evaluation cluster sample). The purpose was to quantitatively evaluate the effect of the criteria decision on the population estimates and/or coverage factors. The difference between the production DSEs and the recomputed DSEs represent the change in population

estimates based on the inclusion of the additional nonmatch cases in followup, assuming the PFU Criteria Evaluation represents truth.

Tables 1 and 2 show the DSEs calculated from the production data within the sample clusters, the DSEs calculated from the PFU Criteria Evaluation results, the differences, and whether or not those differences are significantly different than zero. Data are shown for the site total, tenure, and race/ethnicity. Estimates for age/sex groups were also computed but are not shown in the tables. The DSE population estimates have been corrected using iterative proportional fitting (i.e., raking), and are based on the PES-C estimation method used in the dress rehearsal (Childers, 1998).

No Evaluation PFU Reinterview results were included in these tables. Only PFU Criteria Evaluation changes are included because the PFU Criteria Evaluation sample covered all 187 clusters, whereas the Evaluation PFU Reinterview sample covered a smaller number of clusters. In addition, the main objective is to determine the effect on the estimates by not sending the PFU Criteria Evaluation cases to PFU.

The DSEs shown in the tables, both production and PFU Criteria Evaluation-based, contain data from the 187-cluster evaluation sample and are not equivalent to the official population estimates. These estimates have been weighted to represent the whole sites. Note these DSEs only include people eligible for ICM/PES; group quarters and service-based enumeration areas are not included. Calculations for Menominee are not given because the number of sample clusters in Menominee is small.

Significance was determined at $\alpha = .10$, which is the Census Bureau standard, using the Dunn method of controlling for multiple comparisons. With the Dunn method, the alpha level was divided by the number of comparisons to be made: one for the total, two for tenure, seven for race/ethnicity, and six for age/sex (Toothaker, 1993), to come up with the significance level used in the tests.

As shown in Tables 1 and 2, there were no significant differences in the DSEs calculated using production results versus the PFU Criteria Evaluation results for the 187 Evaluation clusters weighted up to the Census 2000 Dress Rehearsal sites. The results for the age/sex groups were similarly non-significant.

The computation of the DSEs after the PFU Criteria Evaluation interview includes the results of following up on specific cases not followed up in PFU. Since there were no statistically significant differences in the DSEs, there is no reason to believe that sending the PFU Criteria Evaluation cases to PFU affects the DSEs (assuming PES-C estimation methodology). Hence, there is no evidence that the Census Bureau should not use the same criteria to followup people in Census 2000.

Table 1: Comparison of DSEs for Poststrata Marginal Variables, Sacramento

Subgroup	DSEs for Eval Clusters Using Production Results	DSEs for Eval Clusters Using Evaluation Results*	Difference	p-value	Signif
Site Total	403,105 (12,119)	401,483 (12,506)	-1,623 (1,978)	0.41	No
Owner	202,434 (7,968)	202,354 (7,921)	-80 (586)	0.89	No
Renter	200,671 (7,189)	199,129 (7,619)	-1,542 (1,470)	0.29	No
NH White/Other	178,712 (9,195)	177,935 (9,039)	-777 (706)	0.27	No
NH Black	64,025 (3,425)	64,686 (3,649)	661 (395)	0.09	No
NH Amer Ind/Alas Nat	13,156 (541)	13,148 (543)	-8 (57)	0.89	No
NH Native Haw/Pac Isl	2,859 (109)	2,859 (110)	0 (12)	0.97	No
NH Asian	62,291 (1,949)	60,734 (2,780)	-1,557 (1,359)	0.25	No
Hispanic	82,063 (2,904)	82,120 (2,913)	57 (342)	0.87	No

Note: Standard errors are in parenthesis

* Only the PFU Criteria Evaluation results were used in the calculation of the DSEs.

Table 2: Comparison of DSEs for Poststrata Marginal Variables, South Carolina

Subgroup	DSEs for Eval Clusters Using Production Results	DSEs for Eval Clusters Using Evaluation Results*	Difference	p-value	Signif
Site Total	756,533 (46,153)	754,258 (46,104)	-2,275 (2,256)	0.31	No
Owner	554,834 (34,915)	553,362 (36,230)	-1,472 (1,407)	0.30	No
Renter	201,699 (19,039)	200,896 (18,370)	-803 (1,550)	0.60	No
NH White/Other	407,923 (16,047)	405,312 (16,128)	-2,611 (1,968)	0.18	No
NH Black	321,225 (32,824)	321,542 (32,879)	317 (667)	0.64	No
NH Amer Ind/Alas Nat	4,447 (450)	4,452 (451)	4 (10)	0.66	No
NH Native Haw/Pac Isl	435 (44)	436 (44)	0 (1)	0.73	No
NH Asian	7,609 (762)	7,614 (760)	5 (23)	0.83	No
Hispanic	14,893 (1,508)	14,903 (1,506)	10 (41)	0.81	No

Note: Standard errors are in parenthesis

* Only the PFU Criteria Evaluation results were used in the calculation of the DSEs.

The study was conducted by not allowing data to be collected by telephone for half of the eligible cases in the evaluation sample clusters, while attempting to collect the data by telephone for the other half. The phone and personal visit cases were paired as the sample was selected, and the percentage of matches to initial phase people and item nonresponse rates were compared to attempt to measure if there were significant differences by the mode of data collection in our population of interest.

Because of production problems, the sample size for this evaluation is too small to make any strong conclusions, but we found no evidence that the mode of data collection affected the person match rates or the item nonresponse rates.

4. DATA COLLECTION MODE STUDY

This study attempts to measure error due to collecting ICM/PES Person Interview data over the telephone from the interviewer's home using the computer-assisted instrument as opposed to collecting the data using the same instrument during a personal visit.

The ICM/PES Person Interview was CAPI. It was designed to be conducted in person by the interviewer after the completion of the initial phase Nonresponse Followup to avoid contaminating the initial phase data in ICM/PES clusters. However, to alleviate tight schedule demands it was decided to collect data for selected cases by telephone using the (CAPI) instrument before the Nonresponse Followup was finished and ICM/PES personal visits began. The selected cases included those people who responded to the initial phase by mail early in the process and provided a phone number.

5. NET ERROR

The original error profile study design included plans for examining the net effect of a subset of nonsampling error sources. The methodology involved estimating a net nonsampling error and combining it with the sampling, or random, error which occurs because only a sample of blocks (and households) is observed in the ICM/PES (Spencer, et. al., 1998). The subset of nonsampling errors that would have been incorporated into the net effect are as follows: (i) errors in the

collection of data, (ii) matching errors, and (iii) random nonsampling error related to estimation which includes the effects of heterogeneity bias, synthetic estimation error, and ratio-estimator bias.

The planned methodology for computing the net error estimates involved using the results of the EFU in conjunction with the MES to determine more accurate residence status and match codes for everyone who was interviewed in the initial and final phases in evaluation sample blocks on Census Day. This process would be used to obtain a lower bound for net error⁴, which would include the subset of the components of nonsampling error delineated above.

Due to the changes in the ICM/PES matching QA program during production (i.e., 100 percent QA by matching experts) and a predetermined coding specification for a portion of the production workload (i.e., the classification of whole household ICM/PES nonmatching persons from non-proxy interviews as residents), the estimate of matching error based on the MES and the estimate of data collection/instrument error based on EFU are thought to underestimate the actual error. The magnitude of the separate component errors, which feed into the net error estimate, are substantially smaller than what is expected in a national decennial census. Based on the proposed net error estimation methodology, the lower bound for the net nonsampling error would similarly underestimate the true value. Thus, the net error estimates are not included in this paper but will be provided at a later date.

For Census 2000, the net error lower bound, as well as the component error estimates, are not expected to underestimate the true value as much as in the Census 2000 Dress Rehearsal. Regarding the production matching operation, the QA program is not designed to include a 100 percent block cluster rematch by the matching experts. In addition, the 2000 Census Evaluation Followup Interview is currently being redesigned to ensure an accurate and reliable estimate of data collection error.

6. RELATED RESEARCH

⁴ This "lower bound" is not a mathematical lower bound associated with a confidence interval. Instead, this phrase is used to indicate that the net error estimate is a subset of the total net error.

For the 1988 dress rehearsal census of St. Louis and east-central Missouri, Mulry and Spencer offer estimates of the errors of the census, DSE, and undercount estimates. Their 'total error' methodology includes decompositions of error based on the PES into components and summarizing the combined effect of the component errors in a total error estimate (Mulry and Spencer, 1991).

The total error methodology is similar to the original design of Census 2000 Dress Rehearsal Error Profile in the general format in which the error related to the respective post-coverage surveys (i.e., the 1988 PES and 1998 ICM/PES) is described. The Error Profile estimates the magnitude of a subset of individual sources of error and originally intended to incorporate these isolated errors into a net error estimate which serves as a lower bound for total error. The total error model also examined errors individually and attempted to create a combined estimate.

However, the 1988 Dress Rehearsal total error model and the Error Profile differ in at least one critical way, the scope of the study. The total error approach to estimating error in the DSE is to try to identify all the sources of error, estimate their magnitudes, and study their propagation through the estimation process (Mulry and Spencer, 1991). The seven individual components of error included in the total error strategy are model error, P-sample matching error, error in the P-sample reported census day address, P-sample fabrications, error in the measurement of erroneous enumerations, imputation error, and sampling error. The total error model also considers mixed error, that is, error that arises from a mixture of a kind of measurement error known as balancing error and failure of assumptions, but concludes that mixed error is negligible for the 1988 PES.

This comprehensive and ambitious philosophy differs from the Error Profile in that only a subset of nonsampling error is isolated and estimated in this report. The net error portion of the Error Profile was originally designed to incorporate errors in the collection of data (somewhat comparable to the total error component of errors in the measurement of erroneous enumerations), matching error (comparable to the total error P-sample matching error component), and random nonsampling error related to estimation which includes the effects of heterogeneity bias, synthetic estimation error, and ratio-estimator bias. The Error Profile included the study of data collection mode effects, which is not applicable to the 1998 PES.

Other error sources (i.e., fabrication, imputation and sampling error) are not included in the error profile but are examined in separate evaluation reports.

Given this important difference (as well as several others not mentioned here) between the total error model and the Error Profile, 1988 and 1998 dress rehearsal error results should not be compared. The Error Profile was never intended to be a comprehensive, exhaustive delineation of all nonsampling and sampling errors in the ICM/PES, but rather a snap-shot of the major nonsampling errors associated with the final population estimates.

7. REFERENCES

Childers, Danny R. (1998) The Design of the Census 2000 Dress Rehearsal Integrated Coverage Measurement (draft), Bureau of the Census internal memorandum dated May 27, 1998.

Forsman, G. and Schreiner, I. (1991) The Design and Analysis of Reinterview: An Overview. Measurement Errors in Surveys. Ed. Paul P. Beimer et al. New York: John Wiley & Sons, Inc. 279-302.

Mulry, M. H. and Spencer, B. D. (1991) Total Error in PES Estimates of Population. *Journal of the American Statistical Association* 86, 839-863.

Spencer, B. D., Hill, J. M., Haines, D. E. (1998) Accuracy of Block-Level Estimates of Population (preliminary draft), memorandum dated May 31, 1998.

Toothaker, L. (1993) Multiple Comparison Procedures. Ed. Michael S. Lewis-Beck, Newbury Park: Sage Publications.

The 2000 Dress Rehearsal Master Address File Building Process

Lionel Howard, Frank Vitrano, U.S. Census Bureau, Planning, Research, and Evaluation
Lionel Howard, FB 2 Room BH123 Washington, D.C. 20233

Key words: master address file, dress rehearsal building operations

INTRODUCTION

The Census Bureau is developing a nationwide address list called the Master Address File (MAF) to document the address of every living quarters in the United States and will use it to implement the full range of Census Bureau demographic statistical programs, including Census 2000. The MAF building process for the Census 2000 Dress Rehearsal involved a series of operations that built on each other and ultimately resulted in the address list used to conduct the census. The MAF building process differed for areas with mail delivery to predominantly city-style addresses (mailout/mailback areas) and areas with predominantly non-city-style addresses (update/leave areas). City-style addresses are characterized by a house number and street name; non-city-style addresses by rural route or box numbers. This paper summarizes the entire dress rehearsal MAF building process. It should be noted that the data presented provides some insight into what we can expect in the Census 2000 environment, but cannot be generalized to the nation, or compared across dress rehearsal sites.

BACKGROUND

The Census 2000 Dress Rehearsal was conducted in Columbia, South Carolina and the eleven surrounding counties; Menominee County, Wisconsin; and Sacramento, California. Each dress rehearsal site was selected because of its demographic and geographic characteristics, and to provide experience with some of the expected Census 2000 environments. Each site used a different mix of census and statistical procedures.

The *Sacramento, California* site was selected because it contains great diversity among racial and ethnic groups. The *Columbia, South Carolina* site was selected because it contains living situations and socioeconomic characteristics that are not found in a predominately urban environment; the *Menominee County, Wisconsin* site because it includes the Menominee American Indian

Reservation.

The methodology used to evaluate each MAF building operation is specific to the process in which it was conducted. Basic counts and percentages are presented, and in some operations the number of addresses added, corrected, or deleted. Note that the relative impact of each operation could not be fully assessed due to the manner in which data were retained on the MAF extracts.

MAF BUILDING OPERATIONS

The MAF building operations in the dress rehearsal were as follows for the mailout/mailback areas: 1990 Address Control File, May 97 Delivery Sequence File, Targeted Multi-Unit Check, Targeted Canvassing, Postal Validation Check, and Urban Update Enumerate. In the Update/Leave areas: Address Listing, and Update/Leave. In both mailout/mailback and Update/Leave areas the Local Update of Census Addresses and Be Counted/Telephone Questionnaire Assistance operations were conducted.

1990 Address Control File (ACF) and May 1997 Delivery Sequence File (DSF): The ACF and DSF were used to create the initial Master Address File for mailout/mailback areas of the dress rehearsal sites. The ACF is a file of addresses developed by Census, for the 1990 Census, and is based on several initial list operations and a series of coverage improvement operations. The DSF is a file of addresses provided by the United States Postal Service. The two files were matched against each other within ZIP Code and street name.

Targeted Multi-Unit Check (TMUC): TMUC was conducted in Sacramento and the mailout/mailback area of South Carolina. The operation compared the housing unit counts at multi-unit addresses (apartments, rooming houses, etc.) between the 1990 Address Control File (ACF) and the May 1997 Delivery Sequence File (DSF). Where these counts differed, enumerators visited or telephoned (when possible) these basic street addresses, to ensure that the census address list had the correct

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau Publications. This report is released to inform interested parties of research and to encourage discussion.

number of units. Enumerators also updated the unit designations for each unit.

Targeted Canvassing (TC): Targeted Canvassing was conducted in Sacramento and Columbia City, South Carolina. In the operation, local officials were asked to identify and prioritize blocks where they expected hidden housing units to exist. These hidden units were units that post offices may not be aware of because they were in basic street addresses (BSAs) where mail was delivered to one specific place and later distributed to individual units by non-U.S. Postal employees (building managers, landlords, etc.). These BSAs may be recent conversions from single unit addresses (like a basement or garage apartment) or they may be BSAs with purposely hidden units because they are illegal. During canvassing, field staff looked for missing or hidden units in the particular blocks identified by the local officials or a subset of these blocks, depending on how many were identified.

Postal Validation Check: In the Postal Validation Check operation, United States Postal Service employees verified the completeness of the MAF by comparing MAF addresses with the addresses in their carrier delivery routes. The Census Bureau limited the operation to 29 ZIP Codes (seven in South Carolina and twenty-two in Sacramento) that were entirely within the dress rehearsal sites and entirely inside mailout/mailback areas. The operation's primary purpose was to capture late new construction in time for the mail out of census questionnaires.

U.S. Postal Service employees also provided updates to address information for existing units on the MAF. The Census Bureau did not make use of these corrections or the information the U.S. Postal Service provided concerning incorrect or undeliverable addresses, since these preexisting addresses were already in the process for printing and mailing census questionnaires. The U.S. Postal Service provided address adds and deletes at a charge of 17 cent per address.

Address Listing: Address Listing was conducted in Menominee and the Update/Leave areas of South Carolina, and is the initial source for building the MAF in these areas. In the operation, census enumerators went door-to-door to identify the mailing address and physical location of housing units. The enumerators also map-spotted each housing unit on a block map. Enumerators provided a concise description of structures where no address was visible.

Update/Leave: The Update/Leave operation was

conducted just prior to Census day in Menominee and the Update/Leave areas of South Carolina. Enumerators canvassed each block in their assigned area, matching what was found on the ground to the list of addresses in the Update/Leave address register. They updated the register by adding new addresses, deleting addresses they could not locate, and correcting addresses, if necessary. When the enumerator found a new housing unit which was not on the register, they added the unit to the list, map spotted the unit on a block map, and addressed the appropriate form type of questionnaire. The enumerator was also responsible for updating the block map with new street features, corrections to street/road names, and deleting street features that did not exist.

Local Update of Census Addresses (LUCA): The LUCA operation was conducted in all dress rehearsal sites. During the Local Update of Census Addresses operation, local and tribal governments participated, voluntarily, in a partnership program with the Census Bureau to conduct a review of the addresses on the Master Address File. Local and tribal government officials were given the opportunity to review the census address list for accuracy and completeness before the Census Bureau delivered questionnaires.

The Census Bureau gave files of addresses (paper or electronic) and maps (used to identify census geography such as block numbers) to the participating local and tribal governments for their review. The Census Bureau allowed governments one month to review these files and maps, and to provide feedback to Census staff. Participating governments provided feedback in the form of recommended adds, deletes, or corrections of addresses. The Census Bureau then conducted a series of operations to determine whether to accept or reject the recommended actions.

Once the Census Bureau finished processing all of a local or tribal government's suggested changes to the MAF, the agency provided the government feedback identifying which changes were accepted and were rejected. At this stage, the government had the opportunity to review the Census Bureau's results and to provide additional feedback. This was an opportunity for the government to correct information from their previous submissions or to attempt to convince the Census Bureau of the existence of units the agency could not find during LUCA field verification.

After the local government had the opportunity to provide a second round of feedback, the field staff made the final determination about whether to include specific housing units in the census. This step of the process was

known as LUCA Reconciliation. Due to timing constraints, for the most part, the Census Bureau simply accepted any feedback the local or tribal governments gave us at this stage and included all added units in the census process.

It should also be noted that prior to the implementation of the LUCA program, the Census Bureau had experimented with a program called the Program for Address List Supplementation (PALS). In this program, local governments gave the Census Bureau their independent list of addresses. The addresses were compared against the address list maintained by the Census Bureau. Submission of inconsistent and/or nonstandard information, low participation rates, etcetera, led to the cancellation of the program. PALS was only conducted in the Sacramento site. The results are not presented in this paper.

Be Counted and Telephone Questionnaire Assistance (BC/TQA): The Be Counted and Telephone Questionnaire Assistance programs are conducted in both the mailout/mailback and Update/Leave areas of dress rehearsal sites. The operations provide two ways that people can complete a census form if they were not otherwise enumerated.

FINDINGS

Targeted Multi-Unit Check (TMUC)

Of the estimated housing units on the MAF for Sacramento (153,000*), approximately 12.7 percent were canvassed, which are contained within 1,325 basic street addresses. Of the 1,325 BSAs canvassed in the operation, 31.2 percent were resolved by telephone, and 68.8 percent by field verification. The TMUC operation contributed 228 additional housing units to the building of the MAF, and deleted 689.

Of the estimated number of housing units in the South Carolina site** (197,820***), approximately 6.22 percent were canvassed, which are contained within 1,867 basic street addresses. Of the 1,867 BSAs canvassed in the

operation, 15.5 percent were resolved by telephone, and 84.5 percent by field verification. The TMUC operation contributed 274 additional housing units to the building of the MAF, and deleted 1,159.

Targeted Canvassing (TC)

Of the 153,000* estimated housing units in Sacramento, approximately 12.7 percent were in blocks that were canvassed. Targeted Canvassing added 756 units to the MAF. These adds represent a 3.9 percent increase of housing units (HUs) in the blocks canvassed.

Of the estimated number of housing units in the mailout/mailback areas of the South Carolina site** (197,820***), approximately 2.9 percent were canvassed. Targeted Canvassing added 111 units to the MAF. These adds represent a 1.9 percent increase of housing units in the blocks canvassed.

Postal Validation Check

The number of adds paid for by the Census Bureau does not match the number of adds processed in the operations. In Sacramento, 3,189 adds were processed in the operation, however, the Census Bureau only paid for 3,054. In South Carolina, 1,587 adds were paid for, but only 1,223 processed. This difference can be attributed to one or more of the following:

- inconsistent tallying and or/invoicing regarding multiple addresses on a single card,
- the USPS not charging Census for addresses returned very late, and
- the USPS returning addresses so illegible (or incomplete) that they were not keyed.

The Postal Validation Check operation also provided a substantial number of addresses recommended for deletes. Because of the timing of the operation and the

* The actual number of housing units on the MAF in mailout/mailback areas of Sacramento, before this operation, was not possible to determine. Therefore, the number of estimated housing units in the site is used as the base for the percentage of housing units canvassed.

** NO specific housing unit count for Columbia City was available, therefore the estimated housing units in the site were used as the base for the percent canvassed.

***The estimate is based on the estimated number of housing units in the site (252,000), and the percentage of units in the mailout/mailback areas of South Carolina after the final results of the dress rehearsal (78.5%).

inconsistency of how the USPS and the Census Bureau defined a housing unit, we are not able to make use of addresses marked for deletion.

accounted for in the total.

Address Listing

The number of mailing addresses and physical descriptions obtained in the operation are presented below.

Table 1.

PVC Cards Paid For				
	Sacramento		South Carolina	
Adds	3,054	(24.3%)	1,587	(32.7%)
Deletes	9,497	(75.7%)	3,269	(67.3%)
Totals	12,551	(100.0%)	4,856	(100.0%)

A high match rate was also found between new addresses provided by the Postal Service and addresses we already had on the Master Address File.

Table 2.

PVC Match Rate		
	Sacramento	South Carolina
Matched to Units on MAF	1,315 (41.2%)	658 (53.8%)
Did Not Match to Units on MAF	1,874 (58.8%)	565 (46.2%)
Total	3,189	1,223

Lastly, the Geography Division attempted to geocode all of the adds provided by the Postal Service, regardless of whether they matched to units already on the MAF. To the extent possible, computer programs were used to geocode the address. When that was unsuccessful, clerical geocoding was conducted.

Table 3.

Geocoding Results of Adds Not on MAF		
	Sacramento	South Carolina
Computer Geocoded	1,587 (85.0%)	270 (47.8%)
Sent to Clerical Geocoding	281 (15.0%)	295 (52.2%)
Total	1,868†	565

† Six of the 1,874 addresses not matching are not

Table 4.

Units Listed Counts		
	South Carolina	Menominee
Mailing Address	50,595 (75.9%)	1,063 (51.6%)
Physical Description	16,109 (24.1%)	997 (48.4%)
Total	66,704	2,060

Update Leave

The number of added, corrected, and deleted addresses in the operation are presented in Table 5.

Table 5.

Update/Leave Counts		
	South Carolina	Menominee
Added	4,331	96
Corrected	7,543	566
Deleted	4,225	87

Local Update of Census Addresses (LUCA)

In terms of participation rates, the Census Bureau obtained the participation of the City of Sacramento and the Menominee Tribal government in the Local Update of Census Addresses program. In the South Carolina site, 31 of the 60 eligible governments (51.6 percent) participated. These government entities accounted for 98 percent of the 1990 Census housing units in the South Carolina site.

The Local Update of Census Addresses operation varied in the capturing of new addresses, corrections to addresses, and addresses to delete across the sites. The number of initial adds, corrections, and deletes accepted and rejected for the dress rehearsal sites are:

Table 6.

Initial LUCA feedback from Sacramento			
	Adds	Deletes	Corrections
Accepted	155 (5.3%)	0 (0.0%)	3,916 (86.5%)
Rejected	2,763 (94.7%)	0 (0.0%)	612 (13.5%)
Total	2,918	0	4,528

Table 7.

Initial LUCA feedback from South Carolina			
	Adds	Deletes	Corrections
Accepted	3,892 (12.6%)	5,361 (43.2%)	15,187 (56.3%)
Rejected	27,050 (87.4%)	7,053 (56.8%)	11,796 (43.7%)
Total	30,942	12,414	26,983

Table 8.

LUCA feedback from Menominee			
	Adds	Deletes	Corrections
Accepted	25 (100%)	17 (60.7%)	282 (97.6%)
Rejected	0 (0.0%)	11 (39.3%)	7 (2.4%)
Total	25	28	289

It should be noted that there was a large rejection rate of initial submissions across the sites. The Census Bureau rejected some adds because they matched to units already on the MAF or they could not be located in a field verification process. The Census Bureau also rejected some adds because we could not process them due to missing or unreadable information (e.g., no block number). The Census Bureau rejected some deletes and corrections because the local governments did not give the agency the MAF identification number which was needed to identify the referenced unit. The Census Bureau rejected other deletes because they referenced units that were outside of the jurisdiction of their governmental entity.

The total number of added units to the MAF includes adds initially accepted from participating governments,

adds re-added during LUCA reconciliation, and adds added for the first time during the LUCA reconciliation step. Menominee did not provide additional feedback in the LUCA reconciliation step so the changes that the Census Bureau originally accepted are the full extent of the changes the agency made to the MAF based on Menominee's input (See Table 8).

Table 9.

Reconciled Changes to the MAF		
	Sacramento	South Carolina
Added	988	11,621
Deleted	0	5,361
Corrected	3,916	15,187

Be Counted and Telephone Questionnaire Assistance (BC/TQA)

In Sacramento, 1,575 Be Counted questionnaires were received. In South Carolina and Menominee, 783 and 21 Be Counted questionnaires were received, respectively. The total number of addresses added from the questionnaires received are presented in Table 10.

Table 10.

Be Counted/Telephone Questionnaire Assistance Results		
	Questionnaires Received	Addresses Added to the MAF
Sac	1,575	535 (34.0%)
SC	783	521 (66.5%)
Men	21	5 (23.8%)

CONCLUSIONS/RECOMMENDATIONS

As a tool for coverage improvement it was found that Targeted Canvassing was productive in providing adds to the blocks canvassed in Sacramento and South Carolina. However, TMUC found fewer than 300 new housing in both of the sites. It is also unclear from the Dress Rehearsal data whether improving coverage of housing units at multi-unit addresses can be done adequately over the phone. It should be noted that the Targeted Canvassing and TMUC will not be conducted in Census

2000 because we will be doing a 100 percent block canvassing in Mailout/Mailback areas of the country.

The Census Bureau obtained the participation of the City of Sacramento, the Menominee Tribal government, and 51.6 percent of the eligible South Carolina governments in the LUCA program. Although only 51.6 percent of the eligible South Carolina governments participated, they accounted for 98 percent of the 1990 Census housing units in the site. It is recommended that the Bureau continue its efforts to form partnerships with local governments in this coverage improvement operation.

Part of the standard procedure of the U.S. Postal Service (USPS) in the Postal Validation Check operation is to provide address corrections and deletions, in addition to address adds. Because of the timing of the operation and the inconsistency of how the USPS and the Census Bureau define a housing unit, we are only able to make use of new addresses provided in the operation. Also, addresses recommended to be deleted could not be used because existing addresses were already printed on census questionnaires and ready for delivery. However, we pay the postal service for these deletions. In Sacramento, 75.7 percent of the 12,551 addresses we paid for were deletions. In South Carolina, 67.3 percent of the 4,856 addresses we paid for were deletions.

In both Sacramento and South Carolina, there was a high match rate between new addresses provided by the postal service and addresses we already had on the MAF (41.2 percent of the 3,189 in Sacramento and 53.8 percent of the 1,223 in South Carolina). We recommend that, as in the dress rehearsal, systems be put into place to look out for duplicate addresses provided in the Postal Validation Check operation.

In the LUCA operation a substantial number of initial submissions were rejected. The Census Bureau rejected some adds because they matched to units already on the MAF or they could not be located in the field verification operation. They also rejected some adds because the agency could not process them due to missing or unreadable information. Additionally, they rejected some corrections and deletes because local governments did not give the agency the MAF identification number which was needed to identify the referenced unit. As a result, we recommended that the Census Bureau do more to improve the process, and to educate and train LUCA participants to make this coverage improvement operation more efficient. It should be noted that revisions to the LUCA program have been made to improve both the process and training for Census 2000.

In Sacramento, 1,575 questionnaires were received in the Be Counted/TQA operations, contributing 535 new addresses to the MAF. In South Carolina, 783 questionnaires were received contributing 521 new addresses, and in Menominee 21 questionnaires were received contributing 5 new addresses. We recommend that more planning go into the operations of the Be Counted Program including the placement of Be Counted forms in the field and the geocoding of addresses in order to ensure that Be Counted response records have time to make it into the Census process.

The relative impact of each operation on the building of the Master Address File could not be fully assessed in the evaluation. This was largely due to the manner in which data were retained on the Master Address File extracts used in the dress rehearsal. In particular, we could not obtain the universe of addresses going into each operation. The universe of addresses going into each operation would have provided a base against which to measure the relative impact of the operation. Additionally, the Master Address File extracts only retained the results of the most recent field operation. By updating the file with the most recent field operation, it was not possible to determine which operation was the initial input source. With these limitations, we still attempted to gain some sense of each operation's relative impact by examining the summarized information found in a series of MAF extracts.

For Census 2000, we recommended a thorough review of the flags set on the Master Address File that show the relative impact of each operation. In particular, the creation of specific variable fields that follow the history of an address, as they relate to an operation. Fields should also be created to determine if an address was part, or not part, of the initial universe for specific operations. The Decennial Statistical Studies Division is currently working with Geography Division and the Decennial Systems and Contracts Management Office to make these improvements.

Reference

Vitrano, Frank, Howard, Lionel, "Census 2000 Dress Rehearsal Evaluation - An Evaluation of the Master Address File Building Process" July 1999.

PERSON DUPLICATION IN THE CENSUS 2000 DRESS REHEARSAL

John Jones and Danny Childers
U. S. Census Bureau, Washington DC 20233¹

Keywords: Integrated Coverage Measurement, Dual System Estimation

the randomness of E-sample duplication while Section 6 states the final conclusions of the paper.

1. Introduction

Census 2000 procedures were rehearsed in three sites during 1998: Sacramento, California; the Menominee Indian Reservation in Wisconsin; and the Columbia, South Carolina area. In each location, after the Census was taken, an independent enumeration of sampled block clusters was performed for the purpose of census coverage measurement. During the Dress Rehearsal, this process was called Integrated Coverage Measurement (ICM). The people and housing units contained in this independent enumeration is known as the P-sample. People and housing units from the census that are counted in the sampled block clusters are called the E-sample. Both the P-sample and the E-sample contain within sample person and housing unit duplication. This duplication is examined with emphasis on E-sample person duplication.

The data used is derived from the ICM matching at all three sites of the Dress Rehearsal; there are E-sample duplicates at each site and each duplicate has exactly one individual that it duplicates. Individuals that have duplicates are called primary persons. Duplicates are identified by clerical match code during matching and subsequent analyst verification. Person duplication is analyzed by the post-strata gender, race, age, and housing tenure to see if duplication occurs more regularly in one group than in another group and to see if we can accurately predict duplication by examining personal characteristics. We attempt to associate duplication with gender, race, age, and housing tenure and search for significant relationships.

Section 2 discusses the methods used to analyze the data. Section 3 examines the frequency of duplication within various post-strata while Section 4 discusses the percentage agreement between primary persons and duplicates on these same post-strata. Section 5 examines

2. Methodology

The database used consists of all census persons at each site before subsampling. Person duplication at each site is analyzed by four post-strata: gender, age, race, and housing tenure. At each site, there are individual records with missing values of post-strata so that each post-strata variable has a level called missing. The age variable has four non-missing levels (ages 0-17, ages 18-29, ages 30-49, and ages 50 and over). The race variable has two non-missing levels in Menominee and South Carolina. At these sites the levels are White and Other where all nonWhites have been collapsed into the Other category. In Menominee the Other category is primarily American Indian while in South Carolina this Other category is primarily African-American. Sacramento has four non-missing racial categories: White, Black, Asian, and Other where the Other category is primarily Hispanic.

The frequency of duplication is determined by taking the ratio of the number of persons duplicated to the total number of people in sample, both in total and for each level of post strata. These ratios are computed for each site and then converted to percentages. Standard errors of these percentages are calculated using the software package VPLX, which uses replication methods to calculate variances of estimates derived from complex surveys as described in Fay (1990). Once these frequencies and their standard errors are determined, within post-strata comparisons are made at each site to check for significant differences in the frequency of duplication.

These comparisons are made using critical values of t-statistics. These critical values are determined using a multiple comparison of means technique with a Bonferroni adjustment, as described in Hocking (1986). The technique allows the overall type 1 error probability to be .10 for a family of tests at a given site. For example,

¹John Jones and Danny Childers are mathematical statisticians in the Decennial Statistical Studies Division of the U. S. Census Bureau. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official U. S. Census Bureau publications. Research results and conclusions expressed are those of the authors and do not necessarily indicate concurrence by the U. S. Census Bureau. It is released to inform interested parties of current research and to encourage discussion.

the age variable at a given site has four non-missing levels so that pairwise comparison of these levels results in six (4 choose 2) different comparisons. The Bonferroni adjustment reduces the significance level of each individual test so that the overall type 1 error for the entire family of tests is .10. Critical values of t are based on this reduced significance level. Absolute values of observed t statistics are reported and compared with critical values of t . Only comparisons of non-missing levels of post-strata variables are deemed to be important.

Duplicate persons are identified by name and often live in duplicate housing units. To investigate the agreement of duplicate persons with those persons being duplicated (otherwise known as primary persons) on gender, race, age, and housing tenure; we use the subset of the E-sample database consisting of primary persons and their associated duplicates. A new database is formed with each individual record consisting of the primary name, gender, age, race, and tenure linked with the duplicate name, gender, race, and tenure. Some primary persons have more than one duplicate; when this happens a separate record is created for each primary-duplicate pair.

To analyze the randomness of duplication, we make a comparison of percentages within levels of post-strata. For every level of post-strata, we compare its percentage occurrence among the duplicates with its percentage occurrence among the nonduplicates and check for significant differences. Again, standard errors of these percentages are calculated via VPLX. If there are significant differences, then the characteristic (level of post-strata) in question is said to be more (or less) likely to occur among duplicates than among nonduplicates. The characteristic is then not a random occurrence. The critical value calculation and subsequent hypothesis testing proceed in a manner analogous to that used to analyze duplication frequency.

3. The Frequency of E-sample Duplication and a Description of the Duplicates

Table 1 gives the unweighted percentage of duplicates that are in each site and the associated standard errors of these percentages. Menominee has the highest percentage of duplicates but it has the smallest population.

Table 1: Percentage of Duplicates in E-sample

Site	Percentage	Standard error
SC	1.26	0.13
Sac	1.02	0.10
Men	4.22	0.83

Table 2 gives the unweighted percentage of males and females that are duplicated at each site. The percentage that is missing gender is also given. (Standard errors are in parentheses). For each site, only the male-female comparison is made and the critical value for each male-female comparison is 1.65. Neither Menominee ($t=0.49$), South Carolina ($t=0.57$), nor Sacramento ($t=0.85$) exhibit significant gender difference in duplication frequency. While Menominee has person records that are missing gender, none of these individuals happen to be duplicated.

Table 2: Percentage of Gender that are Duplicated

	Male	Female	Missing
SC	1.31(0.14)	1.37(0.14)	0.22(0.10)
Sac	1.11(0.12)	1.03(0.12)	0.32(0.14)
Men	5.21(1.43)	6.05(1.14)	0.00(0.00)

Table 3 gives the duplicate percentage of each age group. There are six comparisons made at each site, implying that the critical value of t at each site is 2.39. Neither South Carolina nor Menominee exhibit significant age difference in duplication frequency. Note that while there are person records with missing age in Menominee, none of these individuals happen to be duplicated. The only significant difference is between persons aged 30-49 and persons aged 50 and over in Sacramento ($t=2.5$).

Table 3: Percentage of Age Grouping that are Duplicated:

Age	SC	Sac	Men
0-17	1.23(0.18)	1.10(0.19)	6.16(1.68)
18-29	1.44(0.24)	1.20(0.20)	5.88(2.22)
30-49	1.37(0.15)	1.20(0.15)	5.36(1.19)
50+	1.45(0.18)	0.84(0.12)	5.40(1.62)
Missing	0.18(0.08)	0.32(0.12)	0.00(0.00)

Table 4 gives duplicate percentage of race in both Menominee and South Carolina while Table 5 gives the duplicate percentage of race in Sacramento (Standard errors are in parentheses). In Menominee and South Carolina only one comparison is made so that the critical value of t is 1.65. In Sacramento six racial comparisons are made so that the critical value of t is 2.39. At each site the observed t value for each comparison is less than the critical value, meaning that there are no racial differences in duplication frequency.

Table 4: Percentage of Race that are Duplicated: South Carolina and Menominee

Site	White	Other	Missing
SC	1.45(0.18)	1.19(0.16)	0.29(0.10)
Men	6.08(2.26)	10.81(7.26)	3.58(0.97)

Table 5: Percentage of Race that are Duplicated: Sacramento

White	Black	Asian	Other	Missing
1.16 (0.15)	1.09 (0.22)	0.78 (0.21)	1.17 (0.20)	0.44 (0.15)

Table 6 gives the percentage of housing unit owners and renters that are duplicates at each site. There is one comparison made at each site so that the critical value of t is 1.65. Menominee ($t=0.20$) and South Carolina ($t=1.18$) have no significant difference in duplication frequency between owners and renters. In Sacramento ($t=2.90$), there is strong evidence that more duplication occurs among renters than among owners. There are person records with missing housing tenure at each site, however, none of those persons are duplicated.

Table 6: Percentage of Housing Tenure that are Duplicated

	Owner	Renter	Missing
SC	1.00(0.21)	1.13(0.29)	0.00(0.00)
Sac	0.80(0.11)	1.37(0.18)	0.00(0.00)
Men	5.92(1.71)	4.69(2.98)	0.00(0.00)

4. Comparison of Characteristics for the Linked Primary-Duplicate Pair

Next we describe the duplicates by examining the extent of agreement on the post-strata between the duplicates and those persons who are duplicated (primary persons). Here, we use the database of linked primary-duplicate pairs. Table 7 gives the percentage agreement between primary persons and duplicate persons at each site. With the exception of age and race in Sacramento and housing tenure in South Carolina, there is at least 89% agreement between primary persons and duplicate persons on these variables.

Table 7: Percentage Agreement between Primary persons and Duplicates on Post-Strata

Post Strata	SC	Sac	Men
Gender	92.9	92.9	97.0
Age	91.1	88.6	94.0
Race	95.8	85.9	89.6
Tenure	78.0	89.9	92.5

Tables 8, 9, and 10 give cross-classifications of age in Sacramento, race in Sacramento, and tenure in South Carolina for the linked primary-duplicate database. These are the three site-variable combinations which have less than 90% percentage agreement. The rows are levels of post-strata for the primary person while the columns are levels of post-strata for the duplicates. Each table shows that the numerous missing data fields contribute most heavily to the low agreement percentage.

Table 8: Cross Classification by Age in Sacramento

	0-17	18-29	30-49	50+	Miss
0-17	103	0	4	0	1
18-29	1	62	4	1	2
30-49	2	4	121	6	4
50+	2	2	5	66	5
Miss	0	0	0	2	0

Table 9: Cross Classification by Race in Sacramento

	White	Black	Asian	Other	Miss
White	174	1	1	6	2
Black	4	47	0	2	0
Asian	0	0	35	4	5
Other	4	0	1	85	12
Miss	4	0	2	3	5

Table 10: Cross Classification by Tenure in South Carolina

	Owner	Renter	Missing
Owner	246	13	65
Renter	14	130	14

5. Randomness of Duplication

To learn which characteristics that duplicates are likely to have, we compare the percentage of the occurrence of that characteristic (level of post-strata) among the duplicate population with that same percentage among the nonduplicate population and check for significant differences.

Table 11 compares the percentage of duplicates that are male and female with the percentage of nonduplicates that are male and female. There are persons with missing gender that are duplicates and persons with missing gender that are nonduplicates but they are not the subject of this study. At each site there are two comparisons made: the percentage of duplicates that are male with the percentage of nonduplicates that are male and the percentage of duplicates that are female with the percentage of nonduplicates that are female. The critical value of the t-statistic is 1.96. For females, the observed

value of t is less than 1.96 in absolute value at each site, meaning that females occur equally among duplicates and nonduplicates. For males, Menominee ($t=1.50$) and South Carolina ($t=0.81$) exhibit no significant difference. However, in Sacramento ($t=2.00$) a significantly larger percentage of duplicates than nonduplicates are male. This does not imply that a significantly smaller percentage of duplicates than nonduplicates are female because of the existence of missing gender fields.

Table 11: Gender Percentage of Duplicates and Nonduplicates

	Female		Male	
	%dup	%ndup	%dup	%ndup
SC	50.81 (4.52)	52.29 (0.49)	42.74 (4.43)	46.27 (0.52)
Sac	47.37 (2.52)	48.58 (0.33)	49.74 (2.40)	44.84 (0.30)
Men	52.24 (6.41)	37.84 (2.64)	46.27 (6.52)	35.81 (2.38)

Table 12 gives the percentage of duplicates and nonduplicates that are in each age category. At each site there are four comparisons made, each comparison is made at an alpha level of .025 and the corresponding critical t value is 2.23. Menominee and Sacramento have no significant differences while South Carolina has a significantly higher percentage of duplicates than nonduplicates over the age of 50 ($t=5.10$).

Table 12: Age group Percentage of Duplicates and Nonduplicates

	Ages 0-17		Ages 18-29	
	%dup	%ndup	%dup	%ndup
SC	22.58 (4.32)	25.27 (1.05)	8.87 (3.27)	14.19 (1.01)
Sac	26.32 (3.75)	24.90 (0.75)	17.11 (2.74)	14.98 (0.46)
Men	31.34 (7.12)	23.13 (3.69)	11.94 (4.54)	8.41 (1.52)
	Ages 30-49		Ages 50+	
	%dup	%ndup	%dup	%ndup
SC	23.39 (4.33)	29.53 (0.64)	37.90 (6.33)	28.22 (1.25)
Sac	32.11 (2.39)	28.45 (0.48)	18.68 (2.55)	23.45 (0.87)
Men	22.39 (4.06)	17.41 (2.11)	32.84 (6.36)	24.11 (2.40)

Table 13 gives the percentage of duplicates and nonduplicates in each racial category for Menominee and South Carolina. At these sites there are two comparisons made so that the critical value of t is 1.96.

Table 14 gives the percentage of duplicates and nonduplicates in each racial category for Sacramento. Here, there are four comparisons made so that the critical value of t is 2.23. There are no significant differences in the occurrence of each race among duplicates and nonduplicates at each site.

Table 13: Racial Percentage of Duplicates and Nonduplicates in South Carolina and Menominee

	Other		White	
	% dup	% ndup	%dup	%ndup
SC	33.87 (8.27)	41.09 (3.09)	64.52 (8.37)	57.21 (3.08)
Men	2.99 (3.21)	2.30 (0.68)	32.84 (9.81)	18.00 (4.23)

Table 14: Racial Percentage of Duplicates and Nonduplicates in Sacramento

	%dup	%ndup
White	46.05 (4.41)	39.32 (1.60)
Black	12.63 (3.15)	13.17 (0.68)
Asian	10.00 (3.03)	15.04 (0.99)
Other	23.16 (3.28)	22.28 (0.87)

Table 15 gives the percentage of duplicates and nonduplicates that are owners and renters. There are two comparisons in each site so that the critical value of t is 1.96. In Menominee and South Carolina there are no significant differences for either owners or renters. In Sacramento, there are significant differences for both owners and renters ($t=2.41$ and $t=3.44$ respectively). A significantly lower percentage of owners and a significantly higher percentage of renters are duplicates at this site.

Table 15: Tenure Percentage of Duplicates and Nonduplicates

	Owners		Renters	
	%dup	%ndup	%dup	%ndup
SC	64.52 (4.74)	59.54 (2.41)	35.48 (4.74)	34.56 (2.43)
Sac	40.53 (4.83)	51.45 (1.82)	59.47 (4.83)	43.68 (1.71)
Men	76.12 (10.96)	53.22 (5.52)	23.88 (10.96)	23.35 (8.48)

6. Conclusions

Duplicates are identified by name, and they generally agree on race, gender, age, and housing tenure. Data capture problems in the form of missing data fields prevent more substantial agreement between primaries and duplicates, although there are examples of disagreement on post-strata.

Duplication occurs among both genders, all races, all age strata, and with both owners and renters. There is a significantly higher percentage of persons aged 30-49 than persons aged 50 and over that are duplicated in Sacramento. However, this does not occur at the other sites. There is a significantly higher percentage of renters that are duplicated than owners that are duplicated in Sacramento. Again, there are no significant tenure

differences at the other sites. Because significant differences do not occur at each site for a given variable, we cannot conclude that there is always more duplication in one level of post-strata than in another level of post-strata.

Similarly, there are examples of significant differences in the percentage occurrence of a characteristic between duplicates and nonduplicates. In Sacramento, a relatively higher percentage of duplicates than nonduplicates are renters and a relatively lower percentage of duplicates than nonduplicates are owners. Also, Sacramento has a relatively higher percentage of duplicates that are male than nonduplicates that are male. South Carolina has a relatively higher percentage of duplicates than nonduplicates over the age of fifty. However, these differences do not repeat themselves at all sites. Therefore, we cannot conclude that there are significant differences in the percentage occurrence of a characteristic (level of post-strata) among duplicates and that same percentage among nonduplicates. The significant differences that do occur are site-specific.

It appears that there is a limited amount to be learned about duplication from examining post-strata alone. We need to investigate the relationship of duplication to census operations to learn about the causes and consequences of duplication.

7. References

Fay, Robert (1990) "VPLX: Variance Estimates for Complex Samples," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Hocking, RR (1986) *Methods and Applications of Linear Models: Regression and the Analysis of Variance* (New York: John Wiley and sons), pp 108-9.

HANDLING OF MISSING DATA IN THE CENSUS 2000 DRESS REHEARSAL INTEGRATED COVERAGE MEASUREMENT SAMPLE

Anne Kearney, and Michael Ikeda, Bureau of the Census*

Anne Kearney, Statistical Research Division, Bureau of the Census, Washington, DC, 20233

Key Words: Noninterview Adjustment, Imputation, Modeling

A. Introduction

This paper outlines procedures used to handle missing data in the Census 2000 Dress Rehearsal Integrated Coverage Measurement (ICM) sample. It also provides a summary of the results of missing data processing. A noninterview adjustment procedure, outlined in Section C, is used to account for whole-household nonresponse. A characteristic imputation procedure, outlined in Section D, is used to assign values for specific missing demographic variables. Finally, persons with unresolved match, residence, or enumeration status have probabilities assigned based on a procedure outlined in Section E. The missing data procedures are generally similar in effect to those used for ICM in the 1996 Community Census and the 1990 Post-Enumeration Survey (PES). Methodologies and analysis of procedures are documented in [1] for the 1990 PES, in [5] for the 1995 ICM, and in [4] for the 1996 ICM. Differences between the Dress Rehearsal ICM missing data procedures and those for 1990, 1995, and 1996 are outlined in [3].

Section B gives some general background. Section F includes results from missing data processing and discussion of their implications. Section G contains conclusions.

B. General Background

The Census 2000 Dress Rehearsal is conducted in three areas: Sacramento, CA; Menominee, WI; and Columbia, SC. The South Carolina site is divided into two subsites for the purposes of ICM sample selection and ICM missing data processing. The ICM sample is selected separately for each site and the two subsites. An overview of the ICM sample design for the Dress Rehearsal can be found in [6]. A general overview of ICM operations in the Dress Rehearsal can be found in [2].

The Dress Rehearsal uses Dual System Estimation (DSE) to calculate estimates. DSE tries to obtain a roster from the ICM blocks independently of the Census. The independent roster (P-Sample) and the Census roster (E-Sample) are matched and the results of the matching are used to estimate the number of persons missed by both

rosters. Estimates are calculated separately for population subgroups called poststrata. Poststratum estimates are summed to marginal totals which are used to calculate the final estimates. The Dress Rehearsal uses a DSE method called PES C. PES C uses person in-movers in the P-Sample poststratum estimates and uses person out-movers to obtain poststratum estimates of match probability for person in-movers. Further details on DSE estimation for the Dress Rehearsal can be found in [7].

C. Noninterview Adjustment

Noninterview adjustment is only performed on the P-Sample. The noninterview adjustment procedure is similar to the procedures used in the 1990 PES and the 1995 and 1996 ICM. However, there are two noninterview adjustments in the Dress Rehearsal because of the use of PES C estimation. The two noninterview adjustments are basically identical to each other, except for the reference date. One noninterview adjustment is based on housing unit status as of Census Day. The other noninterview adjustment is based on housing unit status as of the day of ICM interview. Each noninterview adjustment spreads the weights of noninterviewed units over interviewed units in the same block cluster and similar type of basic address. There are collapsing rules if the number of interviewed units (in the block cluster x type of basic address category) is too small compared to the number of noninterviewed units. Person non-movers and person out-movers are used to determine Census Day housing unit status. Person non-movers and person in-movers are used to determine ICM interview day housing unit status.

Interview: A unit is an interview (for the given reference date) if there is at least one person (with name and at least one demographic characteristic) who possibly or definitely was a resident of the housing unit on the given reference date.

Noninterview: An occupied housing unit (as of the given reference date) that is not an interview is a noninterview.

The noninterview adjustment based on Census Day is used to adjust the weights of person non-movers and person out-movers. The noninterview adjustment based on day of ICM interview is used to adjust the weights of person in-movers.

D. Characteristic Imputation

P-Sample characteristic imputation for the Dress Rehearsal is similar to characteristic imputation for the 1990 PES and the 1996 ICM. In a change from both 1990 and 1996, we use the demographic information from the Dress Rehearsal Census edited file (CEF) for the Dress Rehearsal E-Sample. Edits and imputation are performed on this file. All E-Sample persons matched to the CEF in the Dress Rehearsal. Because of this, no ICM imputation was done in the Dress Rehearsal E-Sample. If we had needed to do ICM imputation in the E-Sample, the methodology would have been basically the same as the P-Sample methodology.

The variables imputed in the Dress Rehearsal are race, Hispanic origin, sex, tenure, and age. P-Sample person mover status is not considered when imputing characteristics. However, persons from a P-Sample whole-household outmover interview are considered to be a separate household for imputation purposes. Age and sex distributions are calculated separately by site.

Tenure is imputed from the previous household with a similar type of basic address (structure code in the E-Sample) with tenure recorded. Missing race is imputed from the distribution of race in the same household. If no one in the household has a nonmissing value of race, then the distribution of the nearest previous household with reported race and similar Hispanic origin is used. Hispanic origin is imputed from the distribution of Hispanic origin in the same household (or the nearest previous household with reported Hispanic origin and similar race if no one in the household has nonmissing Hispanic origin). Age is imputed from the distribution of age for persons with similar relationship to reference person, and age of reference person. For one-person households, age is imputed from the distribution of age in one-person households.

Sex of reference person (with spouse present) or spouse of reference person is imputed by assigning the person with a missing value for sex the sex opposite to that of their spouse. If both reference person and spouse have sex missing, then sex for the reference person is imputed from the distribution of sex for reference persons with spouse present. The spouse is then assigned the sex opposite to that of the reference person. For one-person households, sex is imputed from the distribution of sex in one-person households. For the reference person (with no spouse present) of a multi-person household, the distribution of sex for reference persons of multi-person households with no spouse present is used. For persons (except reference persons and spouses) from multi-person households with non-missing relationship, sex is imputed from the distribution of sex for persons (excluding

reference persons and spouses) from multi-person households. For persons from multi-person households with missing relationship, sex is imputed from the distribution of sex for persons (excluding reference persons) from multi-person households.

E. Assigning Match, Residence, and Correct Enumeration Probabilities

Probabilities for persons with unresolved final Census Day residence (P-Sample), final match (P-Sample), or final correct enumeration (E-Sample) status are estimated by calculating weighted ratios based on persons with resolved final status. Ratios are calculated separately for each site and use the ICM sampling weights. The use of ratios to estimate all three probabilities is new for the Dress Rehearsal. In 1996, hierarchical logistic regression was used to model residence and correct enumeration probability and in 1990 and 1995 hierarchical logistic regression was used to model match and correct enumeration probability.

For Census Day residence status, three separate ratios are calculated. The residence probability for unresolved persons needing followup is the proportion of persons needing followup who are residents. The residence probability for unresolved persons who did not need followup is the proportion of persons not needing followup who are residents. The residence probability for persons with insufficient data for matching is the proportion of all persons who are residents. The proportions are based on person nonmovers and person outmovers with resolved final residence status. The Census Day residence probability for person in-movers is irrelevant to estimation and was set to 0. Note that the residence probability as of the date of ICM interview for person in-movers and person nonmovers is assumed to be 1 (except that infants born after Census Day are not considered to be ICM interview day residents).

Some person nonmovers and person outmovers have unresolved match status. The match probability for these persons is the proportion of matches among person nonmovers and person outmovers with resolved final match status (excluding confirmed Census Day nonresidents). The match probability is set to 0 for confirmed Census Day nonresidents. The match probability for person in-movers is irrelevant to estimation and was set to 0.

For E-Sample persons with unresolved enumeration status, the correct enumeration probability is the proportion of correct enumerations (among persons with resolved enumeration status) in the given match code group. E-Sample match code groups are defined by before-followup match code, whole/partial match code,

address code (HU match status from HU matching), and DSE followup status.

Special Cases

Large clusters were subsampled in the Dress Rehearsal. If an E-Sample person is duplicated with K persons subsampled out of the E-Sample, then the initial correct enumeration probability is multiplied by $1/(K+1)$, since we do not know which person is the "real" person.

A surrounding block search was done in a small number of outlier clusters. Surrounding blocks in Sacramento were generally eligible for Nonresponse Followup (NRFU) and Undeliverable As Addressed vacant (UAA) sampling. If a P-Sample person matched to a surrounding block person from the NRFU or UAA sample, then the match "probability" of the P-Sample person was set equal to the NRFU or UAA weight of the surrounding block person. There were no E-Sample persons duplicated in a surrounding block in the Dress Rehearsal. If an E-Sample person had been verified to belong in a surrounding block and also to be duplicated with a surrounding block person in the NRFU or UAA sample, then the E-Sample correct enumeration "probability" would have been set to one minus the NRFU or UAA weight of the surrounding block person.

F. Results

All counts in this document are unweighted counts. Certain tables display results for only Sacramento. This is in the interest of space and because the results from the other sites were similar.

1. Noninterview Adjustment

Table 1 gives the noninterview rate by site for Census Day interview status and ICM interview day interview status. Noninterview rates based on Census Day status tend to be higher than noninterview rates based on ICM interview day status because all person nonmovers and in-movers (except for persons born after Census Day) are assumed to be ICM interview day residents, while there are residence questions and other operations that can make person nonmovers and out-movers Census Day nonresidents.

Table 1: Noninterview Rates

	Census Day Status		Interview Day Status	
	NI Rate (%)	Occ HU	NI Rate(%)	Occ HU
Sacramento	5.07	15087	2.18	15217
Rural SC	4.04	7377	1.39	7391
Columbia	6.23	8417	2.02	8398
Menominee	1.68	416	0.17	595

NI Rate is the noninterview rate.

Occ HU is the total number of occupied housing units.

2. P-Sample Characteristic Imputation

Table 2 gives the item imputation rates for Sacramento for the five variables that were imputed. Rates are given for three sets of persons. The first set consists of all persons that are included with nonzero weights somewhere in the P-Sample portion of the Dress Rehearsal estimate. Specifically, this includes person nonmovers who are Census Day residents or possible residents from interviewed households based on Census Day interview status, person in-movers from interviewed households based on ICM interview day interview status, and person out-movers from interviewed households based on Census Day interview status. The second set consists of those person in-movers included in the first set. The third set consists of those person out-movers included in the first set.

In general, the imputation rate for in-movers is slightly higher than the overall rate, while the imputation rate for out-movers is substantially higher than the overall rate for age, Hispanic origin, and race. This is probably due to out-mover data often being collected by proxy.

Table 2: Item Imputation Rates (Percent)

Sacramento	All	In-movers	Out-movers
Tenure	0.66	1.39	0.56
Sex	0.41	0.55	1.86
Age	2.14	2.79	8.90
Hispanic Origin	1.29	1.82	11.04
Race	2.13	2.53	13.01
Total persons	37968	2368	1775

3. E-Sample Characteristic Imputation

The variables needed to assign poststrata (tenure, race, Hispanic origin, age, and sex) were obtained from the Census Edited file. Because of this, there was no missing data for these variables and no actual E-Sample imputation was done by the ICM missing data system.

4. Modeling for Unresolved Status

General Overview

Table 3 gives information on the proportion of persons with unresolved status. Note that P-Sample persons with insufficient information for matching are unresolved for both residence status and match status, as are P-Sample persons with a final code of possible match. The proportion of unresolved persons is fairly small. Results from the 1995 ICM [8], [9], [10] suggest that the method for modeling for unresolved status does not have a major

effect on the estimates. Note that there was a substantially higher proportion of unresolved persons in the 1995 ICM, since roughly half of the persons needing followup were sampled out of followup in 1995.

Table 3: Unresolved Status

Note that a few P and E-Sample persons identified as not needing followup have unresolved final status.

a. Percent Unresolved (Overall)

	P-Sample		E-Sample
	UR	Insuff	UR
Sacramento	3.10	1.17	3.64
Rural SC	1.55	0.66	1.47
Columbia	2.51	1.04	3.33
Menominee	1.54	1.43	0.50

P-Sample percentages are percentages of Census Day residents and possible residents from interviewed households (based on Census Day interview status). E-Sample percentages are percentages of E-Sample persons. P-Sample UR refers to persons with unresolved final residence status. E-Sample UR refers to persons with unresolved enumeration status before accounting for duplication with persons subsampled out of the E-Sample. Insuff indicates insufficient information for matching.

b. Persons Sent to DSE Followup

	P-Sample			E-Sample	
	Tot	%UR	%UR M	Tot	%UR
Sacramento	3306	20.7	0.5	5470	21.68
Rural SC	1144	13.1	0.6	2247	10.95
Columbia	1146	19.3	0.4	3205	17.72
Menominee	86	15.1	0.0	101	5.94

P-Sample Tot is the number of residents and possible residents sent to followup. P-Sample percentages are percentages of P-Sample Tot. P-Sample UR are unresolved residents, UR M refers to unresolved match status after followup. E-Sample Tot is the number of E-Sample persons sent to followup. E-Sample percentages are percentages of E-Sample Tot. E-Sample UR are unresolved enumeration status.

P-Sample

We see in Table 4 that DSE followup in the Dress Rehearsal resolved the match status of almost all persons sent to followup in Sacramento. We also see that DSE followup almost never changed a before followup match to a nonmatch (except for before followup matches to surrounding blocks) and rarely changed a before followup nonmatch to a match. Possible matches could become either matches or nonmatches (but more frequently became matches). Note that confirmed

nonresidents are not in the table. DSE followup confirmed 551 persons as nonresidents in Sacramento.

Table 4: Before Followup Match Code and Final Match Code for P-Sample Persons Sent to Followup (Except for Confirmed Nonresidents)

BFU Match Code	Sacramento Final Match Code							Total
	MR	MS	MU	NR	NU	PKP		
Match (M)	202	0	71	0	1	0	0	274
Match Sur BI (MS)	0	2	0	0	9	0	0	11
Nonmatch (NP)	33	0	0	2167	570	1	1	2772
Poss Match (P)	165	0	6	53	9	16	0	249
Total	400	2	77	2220	589	17	1	3306

MR is matched resident, MS is matched resident, matched to person in surrounding block, MU is matched with unresolved residence status, NR is nonmatched resident, NU is nonmatched with unresolved residence status, P is possible match, KP is match not attempted due to incomplete or invalid name

Table 5 shows the estimated residence probabilities assigned to persons with unresolved residence status in each site.

Table 5: Estimated Residence Probabilities

Followup Status	Site			
	Sacramento	Rural SC	Columbia	Menominee
Sent	0.826	0.808	0.742	0.771
Not Sent	0.991	0.988	0.989	0.985*
Insuff Info	0.976	0.976	0.972	0.969

* No unresolved persons in this category

For illustration purposes, Table 6 contains the counts of confirmed residents, confirmed nonresidents, and unresolved persons by match code group for Sacramento. All persons included in the residence probability calculations are included in Table 6. The proportion resident in BFUGP 1 (matches and possible matches sent to followup) tends to be somewhat higher than for other persons sent to followup; the proportion resident in BFUGP 3 (whole household nonmatches) tends to be somewhat lower.

Table 6: Residence Status by Match Code Group

BFUGP*	Sacramento			
	Resident	Nonres	Unresolved	% Resident of Resolved
1	422	18	112	95.91
2	1495	237	284	86.32
3	705	296	288	70.43
4	31779	276	3	99.14
5	0	0	421	---

*BFUGP is the P-Sample Match Code Group. BFUGP 1-3 are sent to followup. BFUGP 1 are matches and possible matches. BFUGP 2 are partial household nonmatches. BFUGP 3 are whole household nonmatches. BFUGP 4 are persons resolved before followup. BFUGP 5 are persons who have insufficient information for matching (before followup status).

Table 7 gives a further breakdown of residence status for BFUGP 3 (whole household nonmatches) for Sacramento. A conflicting household is where the housing unit matched and both the P-Sample and E-Sample collected persons but none of the persons in either the P-Sample or E-Sample households were matches or possible matches. The proportion resident for persons from conflicting households tends to be lower than for the other persons from BFUGP 3.

Table 7: Residence Status for Whole Household Nonmatches

Address Code	Sacramento			
	Resident	Nonres	Unres	% Resid of Resolved
HU Matched	215	29	101	88.11
HU Not Matched	57	6	66	90.48
Conflicting HH	433	261	121	62.39

The "HU Matched" row excludes persons from conflicting households.

Table 8 contains the estimated match probabilities assigned to persons with unresolved match status. Most of the persons with unresolved match status are persons with insufficient information for matching (the others are possible matches).

Table 8: Estimated Match Probabilities

	Sacramento	Rural SC	Columbia	Menominee
Est Match Prob	0.780	0.727	0.843	0.829

Table 9 contains the counts of confirmed matches, confirmed nonmatches, and persons with unresolved match status by person mover status for Sacramento. Most persons with unresolved match status have

insufficient information for matching. Persons in Table 9 are Census Day residents or possible residents. The proportion matched tends to be slightly lower for person outmovers than for person nonmovers.

Table 9: Match Status by Person Mover Status

	Sacramento			
	Match	Nonmatch	Unres	% Match of Resolved
Nonmover	26528	7045	256	79.02
Outmover	901	591	183	60.39

E-Sample

Table 10 shows the estimated initial correct enumeration probabilities assigned to persons with unresolved initial enumeration status in each site. Initial correct enumeration probabilities are later modified to account for duplication with persons subsampled out of the E-Sample. Confirmed erroneous enumerations could have had their probabilities modified for duplication in surrounding blocks if there had been any such duplication.

Table 10: Estimated Initial Correct Enumeration Probabilities

BFUGP**	Site			
	Sacramento	Rural SC	Columbia	Menominee
1	0.940	0.879	0.943	0.475
2	0.874	0.849	0.864	0.774
3	0.741	0.732	0.800	0.384
4	0.848	0.714	0.971	0.897*
5	0.951	0.888	0.959	0.944*
6	0.000*	0.000*	0.000*	0.000*

* No unresolved persons in this category.

**BFUGP is the E-Sample Match Code Group. BFUGP 1-4 are sent to followup. BFUGP 1 are matches and possible matches. BFUGP 2 are partial household nonmatches. BFUGP 3 are whole household nonmatches where the address is matched. BFUGP 4 are whole household nonmatches where the address is not matched. BFUGP 5 are persons resolved before followup. BFUGP 6 are persons with insufficient information for matching (before followup status).

Table 11 gives a further breakdown of residence status for BFUGP 3 (whole household nonmatches where the housing unit matched in housing unit matching) for Sacramento. A conflicting household is where the housing unit matched and both the P-Sample and E-Sample collected persons but none of the persons in either the P-Sample or E-Sample households were matches or possible matches. The proportion correct

among persons from conflicting households in NRFU tend to be lower than for the other persons from BFUGP 3.

Table 11: Initial Correct Enumeration Status for Whole Household Nonmatches Where the HU Matched

	Sacramento			
	Corr	Erron	Unres	% Corr of Resolvd
HU Matched	1089	156	559	87.47
Conflict HH, Not NRFU	137	53	54	72.11
Conflict HH, NRFU	117	248	85	32.05

The "HU Matched" row excludes persons from conflicting households.

G. Conclusions

The Dress Rehearsal ICM Missing Data seems to be generally satisfactory. We probably want to put E-Sample persons from conflicting households in NRFU in their own match code group as these persons seem to have a lower probability of being correct. We may also want to split the remaining persons from BFUGP 3 (whole household nonmatches where HU matched in HU matching) into two match code groups: non-NRFU conflicting households and the remainder. On general principle, we may also want to put matches needing followup and possible matches needing followup into separate match code groups.

For the P-Sample, we may want to calculate P-Sample residence probabilities separately by match code group. We may also want to put P-Sample persons from conflicting households in their own match code group and put matches and possible matches needing followup into separate match code groups. In addition, we may want to calculate match probabilities for insufficient information people separately for movers and nonmovers.

H. References

- [1] Bureau of the Census internal memorandum from G. Diffendal and T. Belin, "Results of Procedures for Handling Noninterviews, Missing Characteristic Data, and Unresolved Enumeration Status in 1990 Census/Post-Enumeration Survey," July 1, 1991.
- [2] Bureau of the Census internal memorandum from D. Childers to M. Ramos, "DSSD Census 2000 Dress Rehearsal Memorandum Series F-DT-2, The Design of the Census 2000 Dress Rehearsal Integrated Coverage Measurement," November 10, 1998.
- [3] M. Ikeda, A. Kearney, and R. Petroni (1998),

"Missing Data Procedures in the Census 2000 Dress Rehearsal Integrated Coverage Measurement Sample," presented at the 1998 meetings of the American Statistical Association (ASA).

[4] M. Ikeda, A. Kearney, and R. Petroni (1998), "Handling of Missing Data in the 1996 Integrated Coverage Measurement," presented at the 1998 ASA Meetings.

[5] M. Ikeda and R. Petroni (1996), "Handling of Missing Data in the 1995 Integrated Coverage Measurement Sample," presented at the 1996 ASA Meetings (a shorter version of this paper appeared in the 1996 Proceedings of the Section on Survey Research Methods, American Statistical Association, 563-568).

[6] Bureau of the Census internal memorandum from D. Kostanich to M. Lynch "DSSD Census 2000 Dress Rehearsal Memorandum Series A-5, Computer Specifications for the Selection of the ICM Sample for the Census 2000 Dress Rehearsal (R. Sands, D. McGrath, and R. Zuwallack, authors)" November 15, 1997.

[7] Bureau of the Census internal memorandum from D. Kostanich to D. Stoudt, "DSSD Census 2000 Dress Rehearsal Memorandum Series A-38, Computer Specifications for ICM Site Level Estimation and Raking for the Census 2000 Dress Rehearsal (E. Schindler, author)," November 17, 1998.

[8] Bureau of the Census internal memorandum from R. Petroni to E. Vacca, "Effect of Different Methods for Calculating Match and Residence Probabilities for the 1995 P-Sample Data, DSSD DSSD 2000 Census Dress Rehearsal Memorandum Series A-23 (M. Ikeda, author)," January 5, 1998.

[9] Bureau of the Census internal memorandum from R. Petroni to E. Vacca, "Effect of Different Methods for Calculating Correct Enumeration Probabilities for the 1995 E-Sample Data, DSSD Census 2000 Dress Rehearsal Memorandum Series A-28 (M. Ikeda, author)," January 5, 1998.

[10] Bureau of the Census internal memorandum from R. Petroni to E. Vacca, "Effect of Using Simple Ratio Methods to Calculate P-Sample Residence Probabilities and E-Sample Correct Enumeration Probabilities for the 1995 Data, DSSD Census 2000 Dress Rehearsal Memorandum Series A-30 (M. Ikeda, author)," January 28, 1998.

* This paper reports the general results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

CONTINUITY AND CHANGE: THE DECENNIAL CENSUS IN THE 21ST CENTURY¹

Juanita Tamayo Lott, Jay Keller, U.S. Census Bureau

Juanita Tamayo Lott, PRED 1002-2, U.S. Census Bureau, Washington, DC 20233

Key Words: Decennial Census

The census of population each decade is the Census Bureau's largest and only constitutionally-mandated program. At the close of the 20th century, there has been much debate and action about the re-engineering of the decennial census. A closer look at censuses over time suggests both incremental and radical changes throughout more than 200 years to improve census taking. It is reasonable to expect that, in the 21st century, the decennial census will encompass both continuity and change, with improvements to the traditional census.

A Traditional Census

Beginning with the first census in 1790, the traditional view of the decennial census was of an enumerator writing down answers to a list of questions from a head of household in order to capture a profile of the U.S. population every decade. With the introduction of the mailout/mailback questionnaire on a trial basis in the 1960 census, a major change occurred to the traditional census, the shift from observer enumeration to self enumeration (with one person providing responses for an entire household). This radical departure was accompanied by the continuity that the population counts remained based almost exclusively on physical enumeration.

The current debate surrounding methodology for Census 2000--that is whether a count primarily based on physical enumeration but supplemented by sampling for traditionally hard-to-enumerate populations is legal, operationally feasible, and statistically defensible--has brought attention to the meaning of a traditional census in terms of the purpose and goals of the decennial census, as well as methodology.

The basic purpose, as stated in Article 1, Section 2 of the Constitution, of the Decennial Census is representation. This purpose has not changed. However the specific forms of representation have changed over the decades. In 1790, there were only three--representation for apportionment of seats to the House of Representatives among the states, representation for taxation, and

representation for military service. Given these various forms of representation, the decennial census is the central part of a broader statistical system designed to produce data needed to implement legislation, assist in decision-making in both industry and government, and help understand changes taking place in our society (Citro and Cohen, 1985: 16).

The initial and continuing goal of the decennial census is a complete and accurate count of the residents of the United States. While this goal remains for the 2000 and 2010 censuses, it has also been modified and expanded by geographic and demographic changes.

In recent decades, complete and accurate counts at other geographical levels were included to serve a variety of federal and other interests. For example, in 1975 Public Law 94-171 was passed which required the Census Bureau for the first time to make available to the states, decennial census data for other than apportionment purposes. Specifically, states could use population counts from the Census Bureau below the national and state levels. The expectation was raised that the Census Bureau would provide population counts for small geographic areas as small as city blocks. Although P.L. 94-171 requires the Census Bureau to furnish decennial census tabulations to the states, the courts have clearly held that the states may use other data sources for redistricting purposes. With this expectation, accuracy of locating persons in the right state was complicated by accurate placement in the correct smallest geographic unit.

This geographical refinement was accompanied by a demographic one with focus on the differential undercount. With slaves and other non-free persons considered as three fifths of a unit for Congressional apportionment, the concept of a differential undercount was embedded in the Constitution and inherent to the first decennial census and succeeding ones (Anderson, 1988: 12). With preparations for Census 2000, the differential undercount continues, not in terms of some segments of the population being counted as less than whole persons, but in terms of being counted less comprehensively than

¹This is an abbreviated version of the September 10, 1999 paper with the same title and authors.

the general population. These undercounted segments have been defined to include historically disadvantaged minority populations (Office of Management and Budget, 1978:19269-19270; Lott, 1998: 31-47).

While extensive evaluations of the 1990 census by the Census Bureau confirmed the persistent and even increasing undercount of minority populations, they also expanded understanding of the hard-to-enumerate populations beyond these discretely defined groups to include persons with characteristics such as irregular household arrangements, irregular housing, little or no knowledge of English (and in some cases illiteracy in any language), fear of government on the part of sample area residents (leading to concealment of information), and missed or erroneously censused housing units (de la Puente, 1993: 2).

Together, the geographical shift from national to sub-national units and the demographic shift from the national population to targeted undercounted small populations have made the goal of a complete and accurate count more challenging. Since 1950 and particularly 1970, the Census Bureau has attempted to meet this challenge with research directed to key data collection operations and methodology directed to coverage improvement.

A Dynamic, Evolving Census: Focus on Methodology

A balance between methodical research and feasible implementation has characterized the historical evolution of decennial census methodology. The Census Bureau's reputation has been built on the collection of original, primary data via census and survey questionnaires. The bureau has developed outstanding competence in such fields as statistics, demography, and social survey methods. It has pioneered significant advances in methods of data collection and analysis.

At the same time, as American society expands and becomes increasingly diverse, the Census Bureau's major focus on primary data collection has become problematic. As early as the 1970s, the strengths and limitations of primary data collection were noted. "...Yet the Bureau's attention has been sharply focused on one research instrument--the census questionnaire--and on one research framework--the various methods of delivering, retrieving, and interpreting completed census forms. Research and evaluation efforts in this context have produced important findings, most notably the sizable enumerator contribution to response variance. But, as the Bureau becomes increasingly concerned with the enumeration of small population subgroups, styles of problem definition and research design that have served it

well in the past may become less productive of manageable solutions to census-taking deficiencies" (Parsons, 1972: 47).

Till now, the major supplement to primary data collection in the decennial census has been coverage improvement programs. The 1970, 1980, and 1990 censuses, using primarily a mailout/mailback questionnaire to collect original data were all supplemented by coverage improvement procedures. In recent decades, the limitations of this mode of data collection were critiqued and evaluated not just by the Census Bureau but by the panels and committees of the National Academy of Sciences and the General Accounting Office (Goldfield, 1992). These assessments concluded the difficulty and improbability of a complete count of the American people by physical enumeration even with extensive coverage improvements.

The criticism of undercoverage in the decennial census must be viewed in perspective. Very few, if any, censuses and surveys command the high response rate of the decennial census. Moreover, they do not have resources like the Census Bureau to examine and evaluate their performance in-depth with respect to statistical and non statistical errors. It is because of its rigorous assessment of the accuracy and comprehensiveness of the decennial count that the Census Bureau is able to pinpoint undercounts by specific population and geographical area. In particular, the 1990 census documented the wide range, complexity, and diversity of the United States population in an era of increased and diverse immigration, growing and more visible interracial and interethnic unions, and a more mobile population. It also conducted the most extensive evaluation of important limitations.

The National Academy of Sciences' report on the design of Census 2000, *Modernizing the U.S. Census*, while critically delineating the shortcomings of the 1990 census, examined and dismissed radical alternatives to the traditional census for 2000. Specifically, the academy panel dismissed these alternatives stating that a national register for the United States was not a feasible replacement for the decennial census; that an administrative records census was not a feasible option for 2000; that a decennial census conducted by the U.S. Postal Service would not be cost effective or an improvement over the conventional census; and finally that a rolling or sample census was not an acceptable replacement for the 2000 decennial census (Edmonston and Schultze, 1995: 13-14). The report's major conclusion was that, "To improve the census results, and especially to reduce the differential undercount, the panel

recommends that the estimates achieved through physical enumeration and sampling for nonresponse be further improved and completed through survey techniques. The census should be designed as an integrated whole, to produce the best single-number count for the resources available. The Census Bureau should publish the procedures used to produce its final counts, as well as an assessment of their accuracy" (Edmonston and Schultze, 1995: 4).

With less than a year before Census 2000, major reliance on use of a physical enumeration via a mailout/mailback questionnaire ensures a traditional census. Furthermore, the Supreme Court ruled in January, 1999 that for Congressional apportionment an unadjusted head count is the sole acceptable method. For apportionment purposes, administrative records and sampling—for nonresponse follow up or to adjust the initial count—to complement or replace a nationwide physical enumeration were dismissed. For other uses of decennial census data, namely state redistricting and allocation of federal resources, the ability to use more than a national physical enumeration remains unresolved. While the Supreme Court decision did not bar adjustment for other than apportionment purposes and current Bureau plans are to adjust for these other purposes, the states are divided on the use of adjusted decennial census counts. It is within this context that the planning for the 2010 census proceeds.

The Decennial Census in the 21st Century

There are two important lessons to be learned from the results of the censuses from 1950 through 1990 and in the unanswered questions still remaining in the planning, implementation, findings and evaluation of Census 2000. First is the need for preserving traditional components of the decennial census that are effective in accomplishing its goal of a complete and accurate population count while reaching out openly and flexibly to new or improved ways in which to capture the ever-growing complexity of demographic and geographic continuity and change. Second is the need to plan an integrated program where the whole is greater than the sum of its parts. Continuity, change and improvements to the traditional census are manifested in the research and development program for FY 2001-2005.

In preparation for the 2010 census, the Census Bureau has recently contracted with the National Academy of Sciences to advise on future methods which can improve upon traditional census measures, with particular attention to the effective and appropriate uses of administrative records and the role of technology, as well

as conventional research methodology. While such an investigation is in keeping with the continuity of multi-year, multi-decade research and testing, plans for the 2010 census are also receptive to flexibility and change. The latter is particularly important given the progress of relatively recent Census Bureau programs including the American Community Survey (ACS), Community Address Updating System (CAUS), and Administrative Records Research (ARR). If and when fully implemented, the American Community Survey will collect annually long form, socio-economic data traditionally collected in the decennial census. The Community Address Updating System will allow for continuing update of the Master Address File throughout the decade. Administrative Records Research will provide a comparative data base. Additionally the 2010 census will be adaptable to the impact of expanding telecommunications on a mailout/mailback questionnaire.

The 2010 planning program is directed towards two integrations. One is the integration of the components of the decennial census program with each other, and the other is the ongoing, throughout-the-decade integration of the 2010 planning program with related Census Bureau substantive programs, including the American Community Survey, the Master Address File, Administrative Records Research, Demographic Analysis and the Coordinated Operations and Methods to Produce Annual Social Statistics (COMPASS), Integrated Information Solutions including the American FactFinder, and with major Census Bureau supportive programs, such as communications and consultations with all stakeholders, including Census Advisory Committees, and customer satisfaction with products and services.

2010 Census Planning and Testing Program, FY 2001 to 2005

The 2010 census program differs from prior decennial censuses by its much earlier start of full scale planning and testing and its decade-long integration with related non-decennial programs. In the Census 2000 cycle, full-scale planning and testing did not begin until 1993. By 1995, when major census tests were conducted, many alternatives to the census process had not benefitted from complete research due to the truncated time available. The Census Bureau had a limited set of performance measures for assessing design alternatives. Some decisions had to be made with incomplete information. Some promising innovations had to be dropped. A fully-funded planning and testing program

that begins in FY 2001 will: 1) allow two additional years before a major mid-decade test of the most promising alternatives, 2) provide the opportunity for a full use of performance measures to assess and compare options, and 3) allow us to pursue innovations that might otherwise have to be dropped.

The objectives of the 2010 census program in the years FY 2001-2005 are to examine and propose alternatives that will:

- ▶ Reduce the cost of the census
- ▶ Improve coverage of the population by reducing the differential undercount
- ▶ Increase mail response rates and reduce respondent burden
- ▶ Improve our ability to gain accurate responses and locate persons geographically
- ▶ Maintain and refine an open process with all stakeholders throughout the decade
- ▶ Spread the cost of data collection more evenly throughout the decade while reducing risk, simplifying logistics, and improving manageability

To achieve these objectives, we will implement the following strategies:

- ▶ Integrate 2010 census development and design with the implementation of the American Community Survey and the development of the Community Address Updating System and Integrated Information Solutions program
- ▶ Continue reengineering the data collection operation, the major component of the census process, through a program of analyzing and testing:

- New Technologies to Improve Mail Response
 - Incentives to Improve Response
 - Administrative records to enumerate nonresponding households, or to use as a primary data collection method on a decennial or more frequent basis

And through the utilization of the:

- Community Address Updating System (CAUS) to improve mailout coverage
 - ACS program to remove the long form from the decennial census

--ACS program to develop local community knowledge and survey and census support systems

- ▶ Develop, test and analyze potential improvements to the operational components of Census 2000 that will be necessary to meet census objectives.

The full program as currently proposed will yield a number of benefits, both tangible and intangible. An early, fully-funded program will allow the Census Bureau to smooth its decennial census staffing requirements, since the presence of personnel early in the program will relieve major hiring requirements at headquarters later in the decade. Similarly, a 2010 census that does not require a long form because of the integration of the census and the American Community Survey (ACS), and has reduced field address and block listing requirements due to a successfully-incorporated ongoing and updated MAF, will have lower dollar requirements in the years 2009 and 2010. Substantial costs and staffing requirements arose in the middle and late parts of the decade to redesign basic operational systems for Census 2000. Preserving core elements of the Census 2000 systems infrastructure, such as the master activity schedule, cost and progress system, payroll system, and recruiting and contracting programs, will obviate the need for recreating these systems later in the decade, reducing both cost and staff requirements, and increasing their effectiveness due to earlier availability in the cycle. Certain potential innovations, such as an administrative records-based census, add to the opportunities for smoothing costs throughout the cycle. The Census Bureau will determine the best organizational strategy through which to manage a major planning and testing program involving the participation of a large staff and divisions cutting across Census Bureau directorates, seeking efficiencies with the current organizational structure and the availability of decennial census personnel as Census 2000 wanes.

We will measure performance and track the realization of benefits through an ongoing evaluation and comparative analysis of new methodologies or operational improvements using the following performance measurement data:

- ▶ total cost of alternatives using census cost models
- ▶ total accuracy of alternatives including the use

- of total error modeling
- value of the data to the Nation when design alternatives yield different content or geographic levels of accuracy
- managerial feasibility of implementing proposed operational alternatives
- acceptability of the design to stakeholders

The 2010 Census is being planned and will be implemented within the context of geographic and demographic continuity and changes which have shaped the conceptual, methodological and operational improvements to the traditional census over 200 years. More than any prior census, the 2010 Census will be defined by a longer planning lead time which allows for a) smooth integration of all major components with each other and b) effective use of resources throughout the decade. Also, the 2010 census will be distinct from previous censuses by its decade-long integration with major substantive and supportive programs of the Census Bureau.

REFERENCES

Anderson, Margo
1988 *The American Census: A Social History*. Yale University Press, New Haven and London

Choldin, Harvey
1994 *Looking for the Last Percent: The Controversy over Census Undercounts*. Rutgers University Press. New Brunswick, New Jersey

Citro, Constance and Michael L. Cohen, editors
1985 *The Bicentennial Census: New Directions for Methodology in 1990*. National Academy Press. Washington, D.C.

Cohen, Michael, Andrew A. White and Keith F. Rust, editors
1999 *Measuring a Changing Nation: Modern Methods for the 2000 Census*. National Academy Press. Washington, D.C.

De la Puente, Manuel
1993 Why Are People Missed or Erroneously Included by the Census: A Summary of Findings from Ethnographic Coverage Reports. Report prepared for the Advisory Committee for the Design of the Year 2000 Census Meeting, March 5, 1993.

Bureau of the Census, U.S. Department of Commerce, Washington, D.C.

Edmonston, Barry and Charles Schultze, editors
1995 *Modernizing the U.S. Census*. National Academy Press. Washington, D.C.

Goldfield, Edwin D.
1992 Review of Studies of the Decennial Census of Population and Housing: 1969-1992. Commissioned paper prepared for the Year 2000 Census Panel Studies, Committee on National Statistics, National Research Council. Washington, D.C.

Lott, Juanita Tamayo
1998 *Asian Americans: From Racial Category to Multiple Identities*, Alta Mira Press. Walnut Creek, CA

National Academy of Sciences
1978 *Counting the People in 1980: An Appraisal of Census Plans*, National Academy of Sciences, Washington, D.C.

Parsons, Carole W., editor
1972 *America's Uncounted People*. Report of the Advisory Committee on Problems of Census Enumeration, National Academy of Sciences, Washington, D.C.

Robinson, J. Gregory
1988 Perspectives on the Completeness of Coverage of Population in the United States Decennial Censuses. Paper presented at the 1988 annual meeting of the Population Association of America, New Orleans, LA.

Steffey, Duane L. And Norman M. Bradburn, editors
1994 *Counting People in the Information Age*. National Academy Press, Washington, D.C.

Accounting for Changes from the 1990 Post Enumeration Survey Methodology in the 2000 Accuracy and Coverage Evaluation Sample Design

Vincent Thomas Mule, Jr., U.S. Bureau of the Census, Washington, D.C. 20233

Abstract: The Accuracy and Coverage Evaluation (A.C.E.) survey will have a different methodology than the 1990 Post-Enumeration Survey (PES). This research was done prior to the Supreme Court ruling when the Integrated Coverage Measurement (ICM) survey was being designed. Since the A.C.E. sample will be a subsample of the ICM design, studying differences between the ICM and PES will address differences between the A.C.E. and the PES and provide information for the A.C.E. survey design. Previous ICM sample design research used data from the PES while not considering these differences. This research focused on accounting for the changes in methodology when simulating coefficients of variation. The sample design and operational differences between the ICM and the PES were the primary changes investigated. While some differences could be accounted, other 1990 conditions are identified that could not. While this design will not be used in 2000, this research investigated how different variance estimations might have affected the simulated reliability. The effect of this design on minority and non-minority estimates is also discussed.

I. Introduction

This paper presents methods to calculate variance estimates and simulate coefficients of variation (CV) for the Integrated Coverage Measurement (ICM) design that are based on 1990 PES data. Previous sample design research assumed 1990 methods instead of reflecting the 2000 ICM methodology. These methods account for differences between the 1990 PES and the ICM. The first method was the differential weighting of the 1990 design. It did not account for the 2000 ICM design where proportional allocation should lead to more equal weighting. Other changed methods are 1) surrounding block search will not be performed for all blocks and 2) the effects of small block cluster¹ weighting. This analysis attempted to account, to some extent, for the methodology changes.

Results are examined for the ICM sample size of 750,000 housing units. This sample had been allocated to the 50 states and the District of Columbia. Within each state, a proportional allocation was planned. This

design would produce efficient direct state total population estimates. However, we wanted to examine the reliability for state subgroup population estimates. This research examined four demographic group estimates in each state. While these were not going to be the estimates produced in 2000, it allowed us to see how this ICM sample design might have affected minority and non-minority estimates.

This analysis presents a variance estimate methodology for accounting for these changes in this 2000 ICM design. Three types of variance estimate methods are examined: direct calculation, synthetic groupings and a mixture of these two methods.

Section II describes the differences in methodology between the PES and ICM. Section III discusses the research methodology used in this analysis. Section IV provides a brief summary.

II. Difference in Methodology between PES and ICM

Changes between the 1990 and 2000 methodologies that we attempted to reflect in the reliability estimates are:

- **Sample Design:** The ICM plan was to conduct a state-based self-weighting design. This will produce more efficient estimates for state total population estimates. This methodology involves removing the effect of differential weighting of the 1990 PES design and replacing it with the self-weighting of the ICM plan.

The author is a mathematical statistician in the Decennial Statistical Studies Division of the US Census Bureau. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. Research results and conclusions expressed are those of the author and do not necessarily indicate concurrence by the Census Bureau. It is released to encourage discussion.

¹Small block clusters have between 0 and 2 housing units.

- Surrounding block search: In 1990, a surrounding block search of 1 to 2 rings was performed for all sampled blocks. In 2000 ICM, the plan was to only perform a 1 ring surrounding block search for 20% of the sampled blocks. An adjustment has been made to compensate for the decrease in surrounding block search.

Since this research uses 1990 Census and PES data, the methodology assumes that certain results from 1990 occur again in 2000. These results are:

- The Master Address File is 99% complete.
- There was a 98% ICM Response Rate. (This was the response rate for the 1990 PES).
- The total Census 2000 estimated undercount is 1.8%.
- 100% accurate data capture in the Census.

If any of the above conditions are not met then the reliability estimates will increase.

One difference between the 1990 and 2000 ICM methodologies that we have not reflected in these estimates is:

- Handling movers: There is a period of time between Census day and when the ICM interview would have been conducted. During that period, people can and do move. In 1990, a procedure known as PES-B accounted for the movers in the estimation. In the current 2000 plan, a procedure known as PES-C will be used². This methodology does not reflect how the difference in handling movers procedures can affect the reliability estimates.

² PES-B matches the census enumerations at the in-mover's census day address to estimate movers. PES-C estimates match rate by matching out-movers but estimates number and characteristics from in-movers.

III. Research Methodology

Step 1: Obtain Direct Variance Estimates from 1990 PES

The direct variance of the Dual System Estimate (DSE) for a Census division/poststrata were calculated. For purposes of this work, each division had $i = 28$ possible poststrata. The 28 poststrata were formed by the cross-classification of 7 age/sex, 2 race/ethnicity (minority, non-minority) and 2 tenure (owner, renter) categories. The variance was calculated by:

$$\text{Var}(\text{DSE}_{i, \text{E-Sample Estimate}}) = E_i^2 \text{Var}(\text{CF}_i)$$

where E_i is the ratio adjusted weighted E-sample size estimate and $\text{Var}(\text{CF}_i)$ is the coverage factor variance for the i th division/poststratum using a jackknife methodology on 1990 PES data. The ratio adjusted weighted E-sample size was used as an estimate of the unadjusted Census count in the above calculation.

The actual Census count should have been used instead. Because of this, Census counts for each division/poststratum state were obtained. Since the DSE variance is a function of the Census count squared, we recalculated the DSE variance with the actual Census count.

The variance of the Dual System Estimate was set equal to DSE variance using the E-Sample estimate times the Census count squared divided by the ratio adjusted weighted E-sample estimate squared.

$$\text{Var}(\text{DSE}_i) = \frac{C_i^2 \text{Var}(\text{DSE}_{i, \text{E-Sample Estimate}})}{E_i^2}$$

where $\text{Var}(\text{DSE}_i)$ is the variance of the Dual System Estimate, C_i is the Census count, E_i is the ratio adjusted weighted E-sample estimate and $\text{Var}(\text{DSE}_{i, \text{E-Sample Estimate}})$ is the DSE variance using the E-sample estimate in the i th poststratum in a division.

Step 2: Obtain Variances for the Four Collapsed Poststrata in a Division

In addition to looking at the reliability of the total state estimate, we wanted to investigate the reliability of

certain groups within a state. The four groups examined were Majority Owners, Majority Renters, Minority Owners and Minority Renters. The Minority groups consisted of Black, Non-Black Hispanic, Asian and Pacific Islander, and American Indians on Reservations. These four groups were formed by collapsing the 28 poststrata from the previous step across the seven Age-Sex levels. There was a significant amount of covariance among the Age-Sex levels that was reflected in the DSE variance of the four collapsed groups. The national correlation between the Age-Sex levels for each of the four collapsed groups was estimated using 1990 PES data.

The four collapsed poststrata DSE Variances were calculated as follows:

$$\text{Var}(DSE_{i*}) = \sum_{a=1}^7 \text{Var}(DSE_{i*,a}) + 2 \sum_{a < b}^7 \sum \text{Cov}(DSE_{i*,a}, DSE_{i*,b})$$

where $\text{Var}(DSE_{i*})$ = DSE Variance for Division/Collapsed Poststratum (4 Levels), $\text{Var}(DSE_{i*,a})$ = DSE Variance for Division/Collapsed Poststratum/Age-Sex Levels (7), $\text{Cov}(DSE_{i*,a}, DSE_{i*,b})$ = DSE Covariance between Age-Sex Levels within a collapsed poststratum for a Division.

The national covariance structure across the age-sex levels was assumed appropriate for a collapsed poststratum. The DSE covariance between age-sex levels was calculated as follows:

$$\text{Cov}(DSE_{i*,a}, DSE_{i*,b}) = \hat{\rho} \sqrt{\text{Var}(DSE_{i*,a}) \text{Var}(DSE_{i*,b})}$$

where $\hat{\rho}$ = Correlation between national age-sex levels for collapsed poststratum based on the 1990 PES.

Step 3: Obtain Variance Component Independent of Sample Size and Weights

The 1990 PES design had differential weights based on the sampling strata. This step factored out of the DSE

variance in a Division/collapsed poststratum, the effect of sample size and the differential weighting of the 1990 PES. This variance component of the DSE in a Division/collapsed poststratum, σ_{i*}^2 , was calculated from the following formula.

$$\sigma_{i*}^2 = \frac{\text{Var}(DSE_{i*})}{\sum_{j=1}^{n_{i*}} w_{i*,j}^2}$$

where $w_{i*,j}$ = the inverse probability of selection in the 1990 PES, n_{i*} = the number of E-sample people in the i^* th Division/poststratum in the 1990 PES.

Step 4: Obtain the Person Sample Sizes by State

For each state, the block cluster sample sizes needed to be converted into people sample sizes. The block cluster is the basic unit of sampling for the A.C.E. The person sample size for each state was estimated by the number of block clusters allocated times the average number of E-sample cases per block cluster in the division in the 1990 PES. This accounted for different densities of people per block cluster across the United States. The state person sample size was then proportionately allocated to the 4 collapsed poststrata based on the 1990 Census population. Let $n_{i*,s}$ be the resulting sample for collapsed poststratum i^* within state s .

Step 5: Estimate Variance for the Allocation by Division/Poststrata

The variance of the DSE for a Division/collapsed poststrata was shown earlier to be a function of the variance component, σ_{i*}^2 , sample size, n_{i*} , and the weights, $w_{i*,j}$. The amount of sample in each Division/collapsed poststratum will change from the 1990 PES to this design. A new estimate of the variance of the DSE for a Division/collapsed poststrata was calculated based on the new sample size and weights.

$$\text{Var}^*(\text{DSE}_{i*,\text{Division}}) = \sigma_{i*}^2 \sum_{s=1}^k \sum_{j=1}^{n_{i*,s}} w_{i*,j,s}^2$$

where $w_{i*,j,s}^*$ = the inverse probability of selection in the ICM self-weighting design and k = the number of states in a division.

The variance of the Coverage Factor for a Division/collapsed poststrata is equal to the variance of the DSE divided by the square of the unadjusted Census Estimate.

$$\text{Var}^*(\text{CF}_{i*,\text{Division}}) = \frac{\text{Var}^*(\text{DSE}_{i*,\text{Division}})}{C_{i*,\text{Division}}^2}$$

Step 6: Estimate Variances and Simulate Coefficients of Variation for States

Three types of variance calculations estimated three types of methods, direct, synthetic and mixed. The variance estimate of the DSE for direct state estimates and synthetic state estimates was calculated based on the proportional allocation of the medium and large block cluster sample. Direct estimates were calculated by only using the sample allocated to the state.

Synthetic estimates were calculated by forming groupings by Census division. A state demographic group coverage factor variance estimate "borrowed strength" by using the division group coverage factor variance estimate. One limitation of this analysis is that while synthetic variances tend to be lower than direct, the bias introduced is unknown.

A mixed estimate was calculated using a direct estimate for some of the collapsed poststrata and a synthetic estimate for the remaining. The variance estimates were adjusted to account for 1) surrounding block search not being performed for all blocks and 2) the effect of small block weighting.

The variance estimates are calculated by summing over the 4 collapsed poststrata (i^*) in each state. Since there was small correlation among the four collapsed

poststrata at a national level in 1990, the covariance among the four collapsed poststrata was ignored.

For Direct State Estimates:

$$\text{Var}(\hat{X}_{s,D}) = \sum_{i=1}^4 n_{i*,s} w_{i*,s}^2 \sigma_{i*}^2 \text{ADJ}_{\text{Surr. Block}} \text{ADJ}_{\text{Small Block}}$$

where $n_{i*,s}^*$ = the sample size (in persons) allocated to state/poststratum for the design,

$w_{i*,s}^*$ = inverse of the probability of selection for the design,

$\text{ADJ}_{\text{Surr. Block}}$ = adjustment for doing surrounding block search in only 20% of the blocks,

$\text{ADJ}_{\text{Small Block}}$ = adjustment for small block weighting effect.

For Synthetic State Estimates:

$$\text{Var}(\hat{X}_{s,M}) = \sum_{i=1}^4 C_{i*,s}^2 \text{Var}^*(\text{CF}_{i*,\text{Division}}) \text{ADJ}_{\text{Surr. Block}} \text{ADJ}_{\text{Small Block}}$$

For Mixed State Estimates:

$$\begin{aligned} \text{Var}(\hat{X}_{s,M}) = & \sum_{i \in \text{Maj.}} n_{i*,s} w_{i*,s}^2 \sigma_{i*}^2 \text{ADJ}_{\text{Surr. Block}} \text{ADJ}_{\text{Small Block}} \\ & + \sum_{i \in \text{Min.}} C_{i*,s}^2 \text{Var}^*(\text{CF}_{i*,\text{Division}}) \text{ADJ}_{\text{Surr. Block}} \text{ADJ}_{\text{Small Block}} \end{aligned}$$

where direct estimates are used for Non-Minority Owners and Non-Minority Renters and synthetic estimates are used for Minority Owners and Minority Renters.

Coefficients of variation can be calculated using the variance methodologies described above and the 1990 DSE estimates.

IV. Summary

The graphs on the following page show simulated coefficients of variation (CV) estimates based on direct, synthetic and mixed variance methods for the 50 states and the District of Columbia. These five graphs compare the simulated CVs for the Total Population,

Non-Minority Owner, Non-Minority Renter, Minority Owner and Minority Renter.

- The direct simulated CVs were all below 0.5%. This reliability would have been needed in order to use these estimates for reapportionment.
- Non-minority direct simulated CVs were all below 0.9% for the direct estimation. Using synthetic estimation lowered the reliability to between 0.1% and 0.4%.
- For Minority estimates, the direct owner simulated CVs were higher than 3.0% for 10 states. For renters, the direct simulated CVs were higher than 3.0% in 21 states. However, the synthetic CVs were only higher than 3.0% for only three states for both owners and renters.

V. Future Research

Future research will involve using this methodology to simulate CVs for possible A.C.E. sample designs. Possible areas for investigation are:

- Simulating 1990 Poststrata, synthetic state and synthetic congressional district CVs. For various A.C.E. sample designs, these simulated CVs can be estimated. This will allow comparisons of the sample design effect on demographic/tenure, state total population and substate area reliabilities.
- Using different synthetic groupings instead of Census divisions. Alternative groupings may have more similar coverage properties while less bias is introduced.

Graph 1

Distribution of State Total Population OVs by Type of Estimator
(750,000 max)



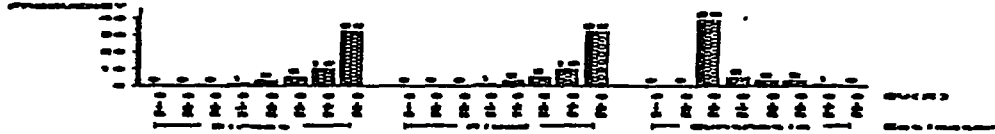
Graph 2

Distribution of State Non-Minority Owner OVs by Type of Estimator
(750,000 max)



Graph 3

Distribution of State Non-Minority Renter OVs by Type of Estimator
(750,000 max)



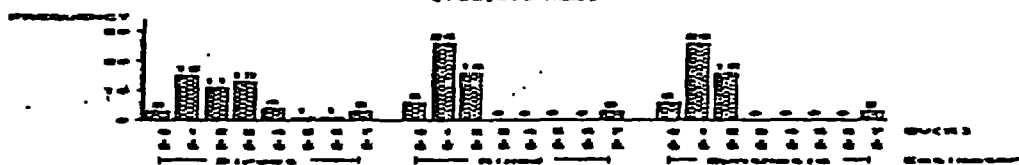
Graph 4

Distribution of State Minority Owner OVs by Type of Estimator
(750,000 max)



Graph 5

Distribution of State Minority Renter OVs by Type of Estimator
(750,000 max)



CENSUS 2000: DEVELOPING A TRADITIONAL CENSUS PLAN

Fay F. Nash, Laureen H. Moyer and Herbert F. Stackhouse, U.S. Census Bureau
Fay F. Nash, U.S. Bureau of the Census, Decennial Management Division, Suitland, MD 20233-7100

Key words: Traditional census, census planning, 2000 Census, coverage improvement

A. Introduction

In the spring of 1998, the United States Census Bureau embarked upon an intensive planning process to develop an alternative approach to conducting Census 2000 without the use of statistical sampling. This plan was publicly released in January 1999. Later that month, the US Supreme Court held that the Census Bureau could not use statistical sampling for reapportionment purposes, but left open the issue of using statistical sampling techniques for other purposes, such as state redistricting, allocation of federal funds and the Bureau's intercensal population estimates program. In response to that decision, the Census Bureau modified the plan for taking the Census by using a more traditional approach, and it is now implementing that modified plan.

This paper first describes the planning process and then discusses the current plan for conducting a traditional census.

B. Focus of the Planning Process

A census without the use of statistical sampling has two fundamental differences: 1) Census Bureau staff would personally visit and collect data from all households not responding to the census by mail and 2) the accuracy of the data must depend solely on improved collection, rather than a combination of data collection and evaluation. Therefore, the planning effort focused on strategies to facilitate conducting the Nonresponse Followup operation and strategies designed to improve quality, in general, and coverage, in particular.

C. Steps of the Planning Process

The planning process began with a review of 1990 Census methodologies and subsequent research to identify programs or operations that were candidates for consideration because of their applicability to the two broad strategies of interest. A compilation of potential operations identified through this process was submitted to the US Congress in April 1998.

An initial framework was developed showing how broad components of programs and operations might fit

together, integrating traditional methods with those components that are necessary for any census methodology. Twenty components were identified for further research by teams of Census Bureau staff; and charters describing the teams' goals and responsibilities were written.

In July 1998 the framework and 20 component charters were in place for consideration by staff teams. All Census Bureau divisions involved in the census were included in the teams which conducted the detailed research and analysis work. For 3 months over 200 Bureau staff participated. This exercise was an enormous drain on staff resources. Census Day was less than two years away and there was intensive work underway on implementation activities applicable to any census design. Additionally, the Bureau was engaged in processing and evaluating data from the 1998 Dress Rehearsal in preparation for Census 2000.

Each team proposed ideas for enhancements to the census plan, analyzing operational issues, evaluating the benefit to be achieved, estimating operational workloads and other cost assumptions, determining activity durations for scheduling purposes, and, finally, recommending those proposals considered by the team to be viable candidates for inclusion in the traditional census plan. The teams consisted of technical staff with expertise in the area of consideration. No limitations of resource needs or cost constraints were placed on the teams as they considered proposals. Therefore the operational analyses conducted were based on more operational and objective measures of data availability, processing capacity, availability of necessary technical knowledge, and the like.

The final step of the planning process was to evaluate each of the recommended proposals, winnow out those that would not be considered further, and to incorporate the resulting proposal into a comprehensive plan that was credible and met the goals for Census 2000. Each proposal was evaluated on several factors. The major factor was its capability to meet the Bureau's data quality goal, defined as reducing the differential and net undercounts while maintaining a minimal level of erroneous enumerations (in which an individual is counted as more than one person). Another important factor considered was the ease of implementation. In the years prior to the Census, the Bureau was restricted by staffing and other budgetary ceilings, both in its regional

offices and at headquarters. Any proposal that required additional experienced staff, was considered difficult to implement (not feasible), given that Census Day was only one year away. Two other factors were considered (though they bore less weight in evaluating and deciding on proposals): impact on public perception and data quality. Cost was not a consideration, provided the proposal was deemed beneficial in improving coverage.

This stage of the planning process was performed by the operational managers from those divisions with responsibility for Census 2000. These were the persons who had the expertise regarding potential coverage improvement gains and other quality measures, and were also the ones with detailed knowledge about the amount of effort required for implementation and availability of resources. The analysis of the proposals and the integration of the final components into a comprehensive plan took considerable time from the operational managers' work schedules during the month of October.

After receiving input from the executive staff of the Census Bureau, the final plan was compiled and made available to the Department of Commerce and then the Congress. Staff also developed a milestone schedule of major activities and a ballpark cost estimate. This cost estimate was a rough approximation because the specific details of each operation had not yet been worked out.

The US Supreme Court decision of January 1999 prompted the Census Bureau to immediately work out the detailed activities of the traditional census plan. The entire 4,000 line Master Activity Schedule was updated to reflect the new plan and details delivered to the US Congress in March. The President's Census 2000 budget request for fiscal year 2000 was revised and delivered to the US Congress by June. The updated plan is being implemented.

D. Traditional Census Components of the Current Census 2000 Operational Plan

There are major operations fundamental to taking a census irrespective of specific design options. For example, the Bureau has to develop a master list of residential addresses and assign those addresses to their proper geography for tabulation purposes. Content must be determined and questionnaires designed and printed to collect the agreed upon data. Field infrastructure and data processing systems need to be designed to collect and process the information. Data dissemination methods need to be established. The Bureau routinely develops an evaluation and research program to inform its planning process for the next decade. This paper deals only with

those components which are unique to the traditional census design or are enhancements to the components planned for the sampling census.

The traditional census components can be categorized into five major classes: 1) methods to improve public response, 2) methods to facilitate the Nonresponse Followup operation, 3) coverage improvement methods, 4) quality assurance and training enhancements, and 5) the accuracy and coverage evaluation program.

1. Methods to Improve Public Response

Methods to improve public response directly affect how much nonresponse followup the Bureau must plan for. Either the US Postal Service or Census staff will deliver questionnaires to most of the addresses in the United States. The Census Bureau projects that 61% of the households will mail these questionnaires back to the Bureau for processing. Thus, census staff need to visit the remaining 39%, or 45 million addresses, to collect the data in person. Needless to say, doorstep interviewing is labor-intensive and requires substantial staffing resources. Strategies that increase public response will directly result in decreasing the Nonresponse Followup workload, facilitating the operation.

The traditional census plan augmented two of the sampling census programs to improve public response: the partnership program, and the paid advertising and promotion program. The partnership program works with governmental entities and community organizations to foster awareness and gain cooperation. The traditional census plan increases the number of partnership staff to 644 so that the Bureau has the ability to foster cooperative relationships with additional governments and community organizations that represent historically hard-to-enumerate groups. The Bureau will also provide limited in-kind funding to support partners in generating awareness of and motivating participation in the census. Examples of such support include printing posters with the community's own logo that promote the Census and buying media time and print space from local media outlets. This is particularly important to small governments and community organizations that lack funds.

Census 2000 is the first census to use a paid advertising campaign rather than to rely on public service announcements to encourage households to return their questionnaires by mail. Motivation to return the census questionnaires is the primary message of the campaign. The traditional census plan expands the paid advertising campaign to include two new messages. The first is an

educational message. Its marketing target is the historically hard-to-enumerate areas of the country. The message is designed to educate the people in these areas about the benefits of the census and why it is important to participate in this major civic endeavor. It will be aired in November 1999 and will continue until the primary advertisement begins in January 2000.

The second paid campaign will run concurrently with the Nonresponse Followup operation. This message is designed to gain public cooperation with the census-takers conducting the personal visit enumeration at addresses for which a questionnaire was not returned. This motivational message is expected to facilitate the enumeration process and may result in improved coverage of persons at those addresses.

The Bureau will also augment its promotional program. Funding is available to expand into non-traditional advertising methods, such as placing posters about the census in community centers, religious centers, and local businesses. The Census Bureau is contracting for a public relations firm to develop a national publicity event that showcases the census throughout the U.S. by involving local participation and media.

The last enhancement under the promotional category is the Census in the Schools program. The Census Bureau has contracted with Scholastic, Inc. to develop a curriculum for elementary and secondary grade students to learn about the census and to use census data in classroom activities. Under the traditional census plan, curriculum materials for the 1999-2000 school year will be available to all teachers in elementary schools, and to all social studies and mathematics teachers in secondary schools. On or about Census Day, teachers will be sent fliers to send home with students describing the census and the importance of the participation of all residents in the census process.

2. Nonresponse Followup

Unlike the sampling census plan, in the traditional census plan, census-takers will follow up on 100 percent of the households that do not return questionnaires by mail. This amounts to a 50 percent increase in workload over what was originally planned. In order to handle this increased workload, the Bureau must hire substantially more temporary workers to conduct the doorstep interviewing and more supervisory staff to supervise the increased enumerator staff. The Bureau estimates that it will employ 860,000 temporary workers during the height of Nonresponse Followup. To provide the infrastructure for the workload increase, the Census Bureau will open

44 additional Local Census Offices, for a total of 520. The time period for conducting the operation was expanded from 6 weeks to 10 weeks to enable the Bureau to contact and enumerate all households.

3. Coverage Improvement Methods

An important aspect of the Census Bureau's plan for reaching all citizen and non-citizen residents is to build a Master Address File using a variety of operations that includes all residential addresses in the United States. In order to account for those housing units that will be built between January 15 and April 1, 2000 (Census Day), the Census 2000 plan includes an invitation for governmental units in mailout/mailback areas to provide a list of newly constructed addresses. The housing units identified by this New Construction operation will be visited during the Coverage Improvement Followup operation in July and August of 2000 (following Nonresponse Followup).

The questionnaire for the traditional census plan has been expanded to allow space for data for up to six residents per household. The sampling census mailback questionnaire only included space for five household members to respond. Since approximately 2.3 percent of households have six members and less than 1.5 percent have more, this improvement will greatly reduce the need for a large household followup. In addition, the quality of self-reported data is considered better than that obtained by an interviewer at a later time.

The largest operational enhancement in Census 2000 is Coverage Improvement Followup. This field operation occurs after Nonresponse Followup and involves the enumeration of several types of housing units. Census staff visit each address designated for followup, determine its status as of Census Day, and obtain the appropriate information on the housing unit and/or its residents. The large majority of the estimated 8 million units that will be visited in Coverage Improvement Followup are those identified during Nonresponse Followup for the first time as vacant or non-existent/deleted. The objective of Coverage Improvement Followup is to improve the census count of persons for housing units that were originally identified incorrectly as vacant or units that were incorrectly designated for deletion from our inventory of housing unit addresses. The Coverage Improvement Followup operation also includes enumeration of housing units with lost or blank mail return questionnaires and of those units identified by local governments during the New Construction operation and late address updates provided by other pre-Census Day operations in 2000.

The traditional census plan includes a Coverage Edit Followup operation by telephone for households that have too many members for the mailback questionnaire. This operation includes a coverage edit operation that is performed by the Bureau to see if there is a discrepancy between the number of persons reported at the beginning of a housing unit's questionnaire and the number for whom data have been collected. If additional persons are identified, census data are collected for them. In addition, the traditional census plan provides for coverage edits to ensure that the correct population count is obtained for all units.

Several other improvements were made in the traditional census plan that will enhance coverage of the population. A variety of special enumeration methods (also called "tool kit" methods because they can essentially be pulled out of a tool kit or reserve of methods that can be used to help fix problems that arise during the Census) will be implemented during Census 2000 in hard-to-enumerate areas. Tool kit enhancements in the traditional census plan include a significantly expanded use of "blitz enumeration" (a compressed time schedule enumeration, usually with a crew of specially trained enumerators) or paired enumeration (when census enumerators work together for safety) to help locate units, to persuade respondents to cooperate, and so forth. These and other targeted operations will be greatly aided by another "tool": the Planning Database that was developed for the traditional census plan. The Planning Database uses 1990 Census data to select sites and neighborhoods where extra efforts are needed for adequate enumeration of the population.

4. Quality Assurance and Training Enhancements

To ensure that major census operations result in high quality data, additional quality assurance measures will be implemented. For example, larger samples of questionnaires will be selected for a reinterview process during the major data collection operation, Nonresponse Followup.

Results of the reinterview will be compared to the original questionnaire data completed by the enumerator. If falsification is identified, appropriate action will be taken against the enumerator and his/her entire work unit will be re-done. Less serious errors will be discussed with the enumerator to prevent such errors in the future. The larger reinterview sample is designed so that all enumerators will have random samples of their work checked. This sample is in addition to the administrative test which targets enumerators whose cases must be rechecked because one or more characteristics of their

work (such as the vacancy rate, the proportion of single person households, and the rate of obtaining data from proxy respondents) is out of tolerance with the work processed from other enumerators in the crew leader district.

Crew leaders and/or their assistants will review all questionnaires completed by the enumerators for completeness. Feedback will be provided as appropriate so that future work from an enumerator reflects the desired level of quality. This review will look for cases where only partial data is collected, refusals that require additional followup, etc. Questionnaires will also be reviewed for potential coverage errors. The crew leader uses a checklist to review all questionnaires to assure that they are complete. Any questions or missing items are returned to the enumerator for resolution. In some cases, the enumerator attempts to determine more information at the housing unit if any reasonable options for doing so remain.

Training will be enhanced for those aspects of the data collection operation that lead to coverage improvement. For example, the enumerator training for Update/Leave operation (the questionnaire delivery method used in more rural parts of the U.S.) will allow more time for developing skills in map reading and recognizing housing units not readily visible from the street.

5. Accuracy and Coverage Evaluation

All modern censuses have included an evaluation survey designed to measure the accuracy of the population data. In Census 2000 the Census Bureau will conduct the Accuracy and Coverage Evaluation Survey (A.C.E.). This survey will consist of approximately 300,000 housing units, or twice the size of the corresponding 1990 Post-Enumeration Survey. This survey will be used to construct dual system estimates of coverage for post strata. Although the state-level census data delivered to the President for apportionment purposes will not reflect results from A.C.E., under the current plan all subsequent data products will. Therefore, data users will have access to the highest quality data for use in redistricting congressional and state legislative districts, analyzing the demographic characteristics of the Nation's population, estimating the population between censuses, and so forth.

E. Current Status

As of this time (August, 1999), the Census Bureau's FY2000 budget is under review by the US Congress. Many of the programs described require substantial funding increases from the level requested for the

sampling plan. The final plan is ultimately dependent on the outcome of the Federal budgetary process. The Census Bureau believes its current plan is the appropriate combination of operations and programs, and that Census 2000 will successfully meet the data needs of the Nation.

End Notes

1. This paper reports the results of planning activities undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of this work and its effect on the Census 2000 plan, and to encourage discussion.

2. The authors wish to thank Carolyn R. Hay and Phillip M. Steel for their careful, professional review and helpful comments regarding this paper.

Bibliography

Census 2000 Operational Plan - Using Traditional Census-taking Methods, Washington, DC: US Department of Commerce, Bureau of the Census, January 1999.

Department of Commerce et al. v. US House of Representatives et al. Case No. 98-404. Washington, DC: US Supreme Ct. Record. Argued Nov. 30, 1998, Decided Jan. 25, 1999.

"New Six Person Mailback Questionnaires." Census 2000 Decision Memorandum No. 62. Washington, DC: October 30, 1998. Unpublished paper.

"Status Report on Planning for a Decennial Census in Year 2000 Without the Use of Scientific Sampling." Washington, DC: US Dept. of Commerce, April 1998.

"Updated Summary: Census 2000 Operational Plan", Washington, DC: US Bureau of the Census, 1999. Unpublished paper.

CONSTRUCTING THE CENSUS 2000 ADDRESS LIST

Robin A. Pennington, Miriam D. Rosenthal, U.S. Census Bureau
Robin A. Pennington, U.S. Census Bureau, Washington, D.C. 20233

KEY WORDS: Master Address File; Housing Inventory; USPS Delivery Sequence File; Housing Unit Coverage; TIGER

Abstract One critical component of producing a complete count of persons and housing units in Census 2000 is a complete address list. In this paper, we review the construction of the Decennial Master Address File, the address list created for Census 2000. The following products and operations will be explained: Master Address File, Decennial Master Address File, United States Postal Service Delivery Sequence File, TIGER, geocoding, Address Listing, Block Canvassing, Local Update of Census Addresses, Mailout/Mailback, Update/Leave, List/Enumerate. The relative importance of each operation will be discussed in regards to the magnitude of the operation and its effect on coverage.

Introduction:

Census 2000 is rapidly approaching. The decennial census is not just a count of people; it is also a count of housing units. People are associated with a geographic area. In order to count people in this fashion, the list of housing units must be accurate. For Census 2000, that housing inventory list is the Decennial Master Address File (DMAF). This is an extract of the addresses on the Master Address File (MAF) maintained by the U.S. Census Bureau. In this paper we discuss the creation of the MAF and the DMAF and their use in Census 2000.

Addresses that are included on the MAF come from a number of sources. Furthermore, the country is divided into areas according to the predominant address type, which determines the enumeration strategy; there are different operations that contribute to the address list in these different areas. The MAF contains a record for every housing unit that has been added to the list in any operation. The status codes from the operations determine whether a Census 2000 questionnaire will be delivered to the housing unit.

Address Sources in Mailout/Mailback Areas:

Files and operations have been chosen for their potential to give a complete listing of housing units, to reflect changes in housing unit status and to yield information

about newly constructed housing units.

The majority of addresses in the country are in what is known for census purposes as the mailout/mailback area, which in general consists of areas with city-style addresses. A city-style address is of the type 121 Main Street. Most of the address list-building operations focus on these areas. The original source of addresses on the MAF for the mailout/mailback areas is the 1990 Census address file, the Address Control File (ACF). The first update to the ACF addresses is a United States Postal Service (USPS) Delivery Sequence File (DSF) of addresses. The USPS updates its DSF every month. The U.S. Census Bureau has procured some of these for the purpose of maintaining a current address list. The DSF lists both residential and non-residential addresses, with an indicator of which type each address is. Address information on the DSF comes from the local mail carriers, so the timeliness of address updates varies by carrier route.

The following paragraphs explain MAF construction, starting from the ACF and DSF address information. The ACF and DSF addresses are processed and assigned Census Bureau geographic codes. This processing occurs on a continuing basis over a number of years. Additional census operations augment and refine the MAF.

Until shortly before the census, the ACF addresses and the DSF residential addresses constitute the MAF. These addresses are tested against Census Bureau geographic information to determine their location at the census block level. The Census Bureau geographic information is maintained in the Topologically Integrated Geographic Encoding Referencing system (TIGER). This is a set of digitized maps for the entire country, as well as for some other regions in which the Bureau performs the census. Digitally connected to the maps for mailout/mailback areas are address ranges for street segments. This database was created around the time of the 1990 Census for the purpose of maintaining consistent and accurate census records.

Within the Bureau's Geography Division, there was an initial operation to assign address ranges to many of the streets in the TIGER files. An address range such as 100-198 is associated in the TIGER files with a street segment, by which is meant one side of a particular street

from one corner to the next. These often correlate with a ZIP+4 designation assigned by the USPS. When the MAF is processed in conjunction with the TIGER files, an address becomes coded to a specific block, where a block is usually represented by a polygon formed by the intersection of streets or streets and natural boundaries.

The coding of these addresses to the geographic level of a block is accomplished in two matching operations. The first is an automated matching operation. When an address on the MAF can be uniquely matched to the address range on a street segment that forms one of the boundaries of a particular block, the address is said to be geocoded to that block.

Addresses for which the appropriate range does not exist in the TIGER files, or for which there is more than one location given in TIGER, are not geocoded during the automated matching. These addresses then enter a second operation, clerical geocoding. Clerical geocoding takes place in all 12 Regional Census Centers. Trained geographic personnel use a variety of reference sources to find the correct location for the addresses and update the TIGER files so that these addresses geocode. Valid and geocoded addresses will appear on each address list used for a field operation.

For some of the addresses, there still is not enough information to locate them correctly in this clerical operation. The MAF indicates no associated block number for such an address. That address will not be included on any field operation address list because a block number must be known; if the housing unit with that address exists, it must be added to the list during the field operation.

There were two DSFs used to obtain addresses before the extant MAF was sent to a 100% Block Canvass field operation. In particular the November 1997 DSF and the September 1998 DSF were incorporated into the MAF. The Block Canvassing operation is the next major address list operation that was undertaken by the Census Bureau for Census 2000. It took place during the months of January - May, 1999 in the mailout/mailback areas of the country. In general housing units on the MAF that have been geocoded to a block are sent out for field verification in Block Canvassing. Added housing units are also anticipated. The Block Canvass listers receive materials for an assignment area, defined geographically. These materials are maps of the numbered blocks in the assignment area and the corresponding list of addresses coded to the specific blocks in the assignment area. Every address in an address register requires an action code from the Block Canvassing lister. The possible

action codes are ✓ for verify; C for a correction to the street name or directional, but not to the house number; D1 for delete; D2 for duplicate, implying the unit exists elsewhere on the list with a different, unmatchable designation such as a different street name or building name; U for uninhabitable; N for nonresidential. There is a Block Canvassing add page for the added housing units. There is also a Block Canvassing Special Place add page for domiciliary situations such as college dormitories and halfway houses. Special Places are enumerated in the census differently from housing units.

One action code that is not allowed due to the block-by-block canvass procedure of the Block Canvassing operation is a block number change for an address that is discovered to be in the wrong block. This can only be accomplished by deleting the unit from the incorrect block with an action code of D1 and adding the unit to the correct block. This may even be performed by two different listers, since the blocks can be in different assignment areas. In order to determine that the address was originally listed in the incorrect block, the addresses must be matched during the processing of the Block Canvassing data. The delete and add will then result in an action referred to as a geographic transfer of the unit.

Occurring in approximately the same time frame as Block Canvassing is a cooperative address list check with local governmental units (GUs) throughout the country. The Local Update of Census Addresses (LUCA) operation occurred in three phases. The phases are distinguished by the enumeration method to be used in the region. For the mailout/mailback areas, the operation is LUCA 1998. The LUCA 1999 operation for non-mailout/mailback areas will be discussed later in this paper. There is also a Supplemental LUCA operation conducted primarily in regions in which the enumeration technique designation changed after the LUCA 1998 areas were identified. This operation is a slight variant on the other LUCA operations and will not be discussed here. In LUCA 1998 the participating GUs received an address list and were asked for input mostly on added units but also on deleted units and corrected street names or directionals. The outcome of this operation is similar to that of Block Canvassing in that units are added to and deleted from blocks, and address corrections are made. Two distinctions between the operations are that the GUs do not necessarily derive the results through a field check but often by such procedures as referring to local address sources and current construction permits. Additionally, because the LUCA 1998 operation focused on changes to the address list, there was no verification of original addresses.

The vintage of DSF that is used for the files sent to LUCA 1998 and to Block Canvassing may differ. This leads to complications in determining the original source of an address as well as the status of an address. As one example of the complications, consider the implications of not incorporating the September 1998 DSF addresses into the list of addresses sent to a LUCA participant compared to incorporating the DSF addresses. The LUCA entity that receives its address list before the incorporation of the September 1998 DSF addresses will have to add any of the addresses that are new to this DSF. The Block Canvassing address list for this region will already have the DSF addresses included. Thus a LUCA 1998 add matches to a September 1998 DSF add, which is then presumably verified in Block Canvassing. The original source of this address on the file is not limited to one operation or file; both the DSF and LUCA 1998 picked this up as a new unit. If the LUCA entity receives an address list after the September 1998 DSF addresses have been incorporated, then LUCA 1998 will take no action on the address and Block Canvassing will verify the address, if it is correct. The DSF is the original source of the address in this case.

Consider the second case, in which both the LUCA entity and the Block Canvassing operation receive the same address list. If there is a new unit that has not yet appeared on the DSF, presumably both the LUCA 1998 operation, which depends on local knowledge, and the Block Canvassing operation, which takes place in the field, will add this unit to its address list. Both the LUCA 1998 and the Block Canvassing operations are credited with adding this address to the address list, and neither can be said to be the sole original source of this address.

A more complicated situation is when the status of a unit does not agree from one operation to another. In the case that LUCA 1998 and Block Canvassing are sent the same list of housing units, LUCA 1998 might add a unit that is not added in Block Canvassing. This will result in LUCA 1998 field verification of the unit. For the case when the addresses from LUCA 1998 are processed in time for inclusion in the Block Canvassing files, a unit added in LUCA 1998 that is deleted in Block Canvassing will result in a field verification of that unit.

If a unit changes residential status from one DSF to the next, the final unit status may be influenced by the timing of the operations. In the case that a unit is residential on the November 1997 DSF but has changed to nonresidential on the September 1998 DSF, a LUCA entity that receives an early address list will have this unit on its list; a LUCA entity receiving an address list incorporating the September 1998 DSF will not have this

unit on its list. It will not be on the address list sent to Block Canvassing either. When an address appears on a LUCA entity's address list, the LUCA participant can either delete the address, code it as nonresidential, make minor corrections or do nothing to it. When it does not appear on the LUCA entity's address list, the LUCA participant can only add the unit. If the unit is readded in either LUCA or Block Canvassing in this scenario, the status conflicts with the DSF status for that unit. The decision for inclusion or exclusion of the unit on the Census 2000 address list must take into account that there are different paths to the same outcome.

If we change the circumstances of the case just described slightly, so that the unit appears as residential on both DSFs, the set of possible final statuses changes, and the set of operation status paths that the unit may travel after such a DSF designation is vastly altered. The unit appears on the LUCA 1998 participant's list. If the LUCA participant denotes the address as a deleted or a nonresidential unit, there will be conflicting information on the address files. The LUCA delete action could match a delete action from Block Canvassing. In the case that the address list from LUCA 1998 is sent to Block Canvassing, the similar situation is that the unit is not readded in Block Canvassing. On the other hand, if a contradictory action is taken in Block Canvassing, then the LUCA 1998 status conflicts with both the DSF designation and the Block Canvassing action.

When there is conflicting information from different operations, there is a hierarchy of operations for the purpose of determining a unit's status. A Block Canvassing action has a higher priority than a LUCA 1998 action, and both Block Canvassing and LUCA 1998 have higher priority than the DSFs, so a unit that receives a delete action in LUCA 1998 but not in Block Canvassing, or is readded in Block Canvassing, will be included on the DMAF. Also from a housing unit coverage standpoint, it is risky to delete a unit from the list when only one operation has determined it to be a delete, particularly when the results of another operation conflict with that designation. In general when only one operation designates an address as non-existent or nonresidential, the address will still appear on the DMAF. If an operation subsequent to the creation of the DMAF designates a second delete, the unit will be flagged on the file as a delete.

Address Sources in Update/Leave Areas:

The initial DMAF was created in July and August of 1999 from criteria based on the action codes given in the operations that preceded its creation. This is the file of

addresses sent for questionnaire printing and labeling. The operations discussed thus far are the operations that occur before the creation of the initial DMAF in mailout/mailback areas. In addition to the operations described above, there are operations that add addresses in non-mailout/mailback areas.

After Mailout/Mailback, the second most common method of questionnaire delivery is Update/Leave. These are denoted as update/leave areas because the address list and maps are updated at the same time that questionnaires are delivered to each housing unit. There are fewer address list-building operations in update/leave areas than in the mailout/mailback areas. The DSF is not used to construct the address list in these regions because the addresses are primarily non-city-style. Instead, the address list for update/leave areas is constructed during a Census Bureau field operation called Address Listing. Census employees are sent to the field with maps of their assignment areas and are instructed to record the city-style address, non-city-style address or location description, or possibly some combination of the above, for every housing unit. In addition the location of the unit is noted on the census map. This operation took place in the fall of 1998.

At the completion of the processing of the Address Listing data, it is possible to tabulate the number of housing units in each block. Because the housing units in these areas may have nonstandard mailing addresses and may be recorded in census files solely with a location description, the GUs participating in the LUCA 1999 operation in these areas are sent lists of housing unit counts by block. When the LUCA 1999 participant disagrees with a Census block count, that block is sent out for LUCA 1999 recanvassing, in which census employees are redeployed to make updates to the address list. After processing the LUCA 1999 recanvassing materials, the block counts are retabulated.

In both the LUCA 1998 and LUCA 1999 operations, there is an appeal process for settling housing unit status or housing unit count discrepancies that are not resolved by the field verification process. Appeals will be resolved after the creation of the initial DMAF.

The operations described thus far yield the list of addresses sent to the contractors for the printing of the Census 2000 questionnaire address labels.

DMAF Updates:

There are a number of update operations that follow the creation of the initial DMAF. These updates to the

DMAF occur when addresses are added in operations up to Census Day, April 1, 2000. There will be November 1999, February 2000 and April 2000 DSFs adding addresses to the decennial census address files. This is an attempt to add newly constructed housing units to the list. The February 2000 DSF will contain the results of a concerted effort on the part of the USPS to update their files, called edit book week. The automated and clerical geocoding operations will take place on these address lists in Geography Division as before. These addresses will need to undergo special processing procedures in order for the housing units to receive questionnaires.

Another address update operation that occurs subsequent to the creation of the initial DMAF is the LUCA 1998 field verification and appeal process. As discussed above, many of the units receiving a conflicting status from the Block Canvassing and the LUCA 1998 operation will be sent for field verification by the Census Bureau; the results of the field verification will be sent to the GUs. At this stage it is possible for the GU to contest the Census Bureau's findings for particular units. At an appeal, the Census Bureau and the GU will submit their evidence of the status of a housing unit for independent review, and a ruling will be issued. Both the field verification and the appeal processes have the potential to change the status of a housing unit.

The last operation in mailout/mailback areas that adds addresses before Census Day is the New Construction operation, another cooperative effort with participating GUs. This operation uses the GU's local knowledge to identify new housing up until Census Day and takes place in February and March of 2000. Addresses added in this operation will also require special procedures for questionnaire delivery.

The last address list-building operation in the update/leave areas is the Update/Leave operation itself. This operation is responsible for having a census questionnaire hand delivered at every housing unit. In the process the MAF and the maps will be updated.

Additional Enumeration Areas:

In the most remote regions of the country, the housing units will be listed at the time of Census 2000. People will be enumerated concurrently. These operations are called List/Enumerate and Remote Alaska enumeration. This will be the only source of addresses in these regions.

Additionally there are special enumeration techniques for some regions of the country in which addresses were listed in previous operations. For example there is an

Urban Update/Leave operation for areas where mail delivery is considered to be problematic. The addresses have passed through all the operations of the mailout/mailback areas up until the time of the census, but the area will be visited by enumerators during the census, and, therefore, additions, deletions and corrections to the address list can be made.

Conclusion:

Ultimately the information from each operation will be fed back to the DMAF so that each housing unit record will contain a complete history of the actions taken in each operation. It will be possible to determine which operation(s) added the unit to the list. The primary operations responsible for adding addresses are the 1990 ACF, the November 1997 DSF, the September 1998 DSF, the November 1999 DSF, the February 2000 DSF, Block Canvassing, Address Listing, New Construction, Update/Leave, List/Enumerate and Remote Alaska, and LUCA 1998, LUCA 1999 and Supplemental LUCA. Because of the timing of events and the overlap of some of the operations, as well as the varying magnitude of the operations, it will not be possible to compare the operations for relative effectiveness in terms of numbers of addresses added to the MAF or deleted from the MAF or of numbers of corrections made to addresses. However with an understanding of the interrelatedness of the operations, some interpretation of the effectiveness of the operations may be attempted.

Disclaimer:

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau Publications. This report is released to inform interested parties of research and to encourage discussion.

THE EVALUATION OF THE BE COUNTED PROGRAM IN THE CENSUS 2000 DRESS REHEARSAL

Dave Phelps U.S. Bureau of the Census, Karen Owens U.S. Bureau of the Census, Mike Tenebaum U.S. Bureau of the Census

Dave Phelps Room BH-119 SFB-2 U.S. Census Bureau Washington, DC 20233

KEYWORDS: Item Nonresponse, Demographics, Be Counted Program

BACKGROUND

The Census 2000 Dress Rehearsal was the culmination of the Census 2000 testing program which began shortly after the 1990 Census was completed. The Dress Rehearsal was conducted in Columbia, South Carolina and eleven surrounding counties; Menominee County, Wisconsin; and Sacramento, California. Each dress rehearsal site was selected because of its demographic and geographic characteristics to provide experience with some of the expected Census 2000 environments. Each site used a different mix of census and statistical procedures. The dress rehearsal provided information to assess procedures used in the individual sites but not for comparisons between sites. One of the goals of the Dress Rehearsal was to measure the effectiveness of making census questionnaires widely available. The Be Counted Program was one means for people to be included in the Census 2000 Dress Rehearsal. This program allowed people who may not have received a Census questionnaire, believed they were not included on a questionnaire, or had no usual residence on Census day the opportunity to pick up and return a Be Counted form.

The Be Counted questionnaires were printed in six languages: English, Spanish, Chinese, Vietnamese, Mien, and Russian. They were accessible in targeted locations that were determined through communications between local census office officials and community partnership specialists. These locations included businesses, community organizations, churches, Department of Motor Vehicle offices, Libraries, Post Offices, and Questionnaire Assistance Centers. There were 218 locations in Sacramento, California, 183 locations in Columbia and surrounding counties of South Carolina, and 16 locations on the Menominee American Indian reservation in Wisconsin (see table 1 for a complete distribution of locations by type). The forms were made available shortly after census day and collected before the start of the operation to personally enumerate households that did not complete a form. Approximately three percent of the 1,700 English forms available at Be Counted locations were picked up by the public in Menominee; about 18 percent of the 59,272 English and Spanish forms were picked up in South Carolina; and

about 39 percent of the 24,249 forms in all languages that were distributed in Sacramento were picked up.

Besides sending in a Be Counted Form (BCF), the public could call the Telephone Questionnaire Assistance (TQA) operation if they believed that they were not counted. This operation had computer instruments that were programmed to help interviewers take census responses from these callers. If the caller did not give the 22 digit identification number off the Census form sent to them, then their response was treated as a Be Counted Form Equivalent. The acronym BCFE will refer to Be Counted forms and TQA responses without the 22 digit identification number.

BCFEs were sent to the Census Geography Division to be geocoded and matched to the Master Address File (MAF). The MAF is a computer address data base that the Census Bureau created, updates, and uses for address information. Geocoding is the process that changes a unique address into its unique identification number.

Those BCFEs that went to Census Geography and contained sufficient address information were geocoded and matched to the MAF. Those addresses that matched to the MAF were assigned the corresponding MAF identification number (ID). Those addresses that did not match to the MAF were sent out to be field verified and if found to be valid were added to the MAF and assigned an ID; if found to be invalid, the BCFE was removed from further Dress Rehearsal processing. Those BCFEs that were returned without sufficient address information to be geocoded were removed from further census processing.

CENSUS GEOGRAPHY RESULTS

The following is a summary for those BCFEs that were sent to Census Geography. Due to a time constraint and late arriving BCFEs some of the BCFEs could not be processed in time to be included in the Dress Rehearsal. Forms that indicated that the person had no address on census day were processed through the Service-Based Enumeration process. Additionally, the procedures for the accounting of BCFEs during processing could not provide trustworthy numbers to report where BCFEs fell out of the process. The following numbers represent those BCFEs that the Census Bureau could account for.

A total of 21 responses were received from Menominee. Of these, eight (38%) did not contain sufficient address information and were removed from further processing, ten (48%) had geocodable addresses that arrived in time for Census processing, and five (24% of all responses) of the geocoded returns were included in the Dress Rehearsal. These five returns contained information for 16 persons who were enumerated in the Dress Rehearsal. There were three (14%) forms that were geocoded, but arrived too late for Dress Rehearsal processing and are not included in the above geocoded counts.

The South Carolina site generated 783 BCFE responses. Of these, 122 (15%) did not contain sufficient address information and were removed from further processing, 606 (78%) had geocodable addresses, and 337 (43% of all responses) of the geocoded returns were included in the Dress Rehearsal. These 337 returns contained information for 821 persons who were enumerated in the Dress Rehearsal. There were 55 (8%) forms that were geocoded, but arrived too late for Dress Rehearsal processing and are not included in the above geocoded counts. (See table 2)

The Sacramento, California site generated 1,575 BCFE responses; 247 (17%) did not contain sufficient address information and were removed from further processing, 907 (57%) had geocodable addresses, and 343 (22% of all responses) of the geocoded returns were included in the Dress Rehearsal. These 343 returns contained information for 870 persons who were enumerated in the Dress Rehearsal. There were 421 (26%) forms that were geocoded, but arrived too late for Dress Rehearsal processing and are not included in the above geocoded counts.

POPULATION RESULTS

Across all three sites, a total of 1,707 persons with address information from BCFEs were included in the Census 2000 Dress Rehearsal. These BCFE's persons were enumerated on four of the eight different form types that make up BCFEs. The four form types included the following: both long and short form versions of TQA enumerations, English Be Counted forms, and Spanish Be Counted forms (see table 5 for distribution of persons included by form type). After processing no enumerations were included in the Dress Rehearsal from the Be Counted form types in Chinese, Vietnamese, Mien, or Russian (a total of 192 forms, see table 4).

To see if there was a difference in the demographic variables of sex, age, race, and Hispanic origin between persons enumerated on BCFEs and persons enumerated

on other mail returns, a chi-square test was calculated on the two populations consisting of Be Counted persons and other mail return persons ($\alpha=0.1$). The Be Counted population includes all BCFE persons that were enumerated in the Dress Rehearsal and the other mail return population includes all persons that were enumerated on other mailed back returns. These distributions are based on self reported data, therefore item nonresponse will change the total population counts across variables in table's 6 and 7. The Menominee site had a total of 16 persons, so many of the cells in a chi-square test had expected values that were too low, hence this comparison is not reported for Menominee.

The results showed a significant difference in the distribution of age, race, and Hispanic origin between the two populations in Sacramento. Persons enumerated on BCFEs were more likely to be either younger, Hispanic, or non-white when compared to those enumerated by other mail returns (see table 6). The results also showed a significant difference in the distribution of age and race between the two populations in South Carolina. Persons enumerated on BCFEs were more likely to be either between the ages of 5 to 14 or 65 and over, or to be Black or African American than they were on the other mail returns (see table 8).

ITEM NONRESPONSE RESULTS

We next examined data quality as measured by item nonresponse rates. Only persons enumerated on BCF mail returns were included in the analysis of nonresponse rates. Due to the small number of BCF persons in Menominee, this site was not included in the analysis of item nonresponse. The TQA returns were excluded since the nature of a Computer Assisted Telephone Interview instrument lowers item nonresponse rates.

The item nonresponse rates for BCF persons were compared to the nonresponse rates of persons enumerated by other mail returns. A chi square test was used to determine if there were any significant differences ($\alpha=0.1$). The variables of interest were sex, age, race, and Hispanic origin. The results for Sacramento indicated that there was a significant difference in item nonresponse rates between the BCF population and the other mail population for all four variables (see table 9). Each of these variables had a higher nonresponse rate for the BCF population than for the other mail return population. The results for South Carolina indicated that there was a significant difference in item nonresponse rate for the Hispanic origin variable (see table 7). This was higher for the BCF population than for the other mail return population. This is a concern as data must be imputed

when answers to requested items are missing, since there is no followup to obtain the information.

RECOMMENDATIONS

A number of recommendations for Census 2000 are made as a result of this evaluation. There are five recommendations:

- Improve the accounting for and documentation of the flow of Be Counted responses through all operational processes, including check-in, geocoding, and field verification as this will assist the Census Bureau in determining exactly where and why Be Counted forms are removed from the processing flow.
- Part of the success of the Be Counted operation is determined by the quality of the data received on forms. Analysis shows that item nonresponse rates are significantly higher on BCFs that they are on other mail returns. An evaluation should be conducted to determine the benefits of including Be Counted responses in a telephone follow-up operation to improve content.
- More planning should go into the operations of the Be Counted Program including the

placement of Be Counted forms in the field and the geocoding of addresses in order to ensure that Be Counted response records have time to make it into the Census process.

Field observers reported that people often had trouble finding Be Counted forms in places where they should have been. The Census Bureau should choose better targeted sites and increase notices and posters in sites to help alleviate these complaints and increase access to the forms.

Conduct additional research to gain insight into the need for and effectiveness of foreign language Be Counted forms, since all of the foreign language forms except some Spanish forms were removed from processing.

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau Publications. This report is released to inform interested parties of research and to encourage discussion.

Table 1: Number of BCF Distribution Centers by Type of Location and Site

	Sacramento	South Carolina	Menominee
Business	136	79	4
Community Organization	8	16	6
Church	0	8	0
Department of Motor Vehicles	2	1	0
Library	6	16	0
Post Office	9	9	2
Questionnaire Assistance Center	52	47	4
Other	5	7	0
Total	218	183	16

Table 2: BCFEs Received in Geography, Geocoded, and Included in the Dress Rehearsal for Menominee

	Received	Geocoded	Included in Dress Rehearsal
BCF-English	20	10	5
TQA	1	0	0
Total	21 (100 %)	10 (47.6 %)	5 (23.8 %)

Table 3: BCFEs Received in Geography, Geocoded, and Included in the Dress Rehearsal for South Carolina

	Received	Geocoded	Included in Dress Rehearsal
BCF-English	548	411	247
BCF-Spanish	2	1	0
TQA	233	194	90
Total	783 (100 %)	606 (77.3 %)	337 (43.0 %)

Table 4: BCFEs Received in Geography, Geocoded, and Included in the Dress Rehearsal for Sacramento

	Received	Geocoded	Included in Dress Rehearsal
BCF-English	984	578	282
BCF-Spanish	173	63	44
BCF-Cantonese	32	28	0
BCF-Mein	2	1	0
BCF-Vietnamese	82	70	0
BCF-Russian	31	24	0
TQA	271	143	17
Total	1,575 (100 %)	907 (57.6 %)	343 (21.8 %)

Table 5: BCFE Persons Included in the Dress Rehearsal by Form Type

	Sacramento	South Carolina	Menominee
BCF-English	707	625	16
BCF-Spanish	129	0	0
TQA, Short	28	156	0
TQA, Long	6	40	0
Total	870	821	16

Table 6: Demographic Comparison of BCFE and Other Mail Return Populations in Sacramento

	Be Counted	Other Mail	
	%	%	p value
Sex	N=838	N=224,477	0.724
Male	46.18	46.79	
Female	53.82	53.21	
Age	N=778	N=222,337	0.001
Under 5	8.61	5.11	
5 to 14	21.08	12.93	
15 to 24	15.17	12.00	
25 to 44	29.95	30.70	
45 to 64	16.97	23.36	
65 and over	8.23	15.89	
Race	N=728	N=207,281	0.001
White	32.55	59.52	
Black or African American	19.92	13.03	
American Indian and Alaskan Native	2.61	1.20	
Asian	30.08	16.75	
Native Hawaiian and Other Pacific Islander	3.43	0.92	
Some other race	7.14	5.27	
Two or more races	4.26	3.31	
Hispanic Origin	N=719	N=215,042	0.001
Yes	33.10	18.89	
No	66.90	81.89	

Table 7: South Carolina Item Nonresponse for 100% Person Data Items

	Be Counted	Other Mail	
	N=625	N=397,308	p value
Sex	1.92	1.19	0.092
Age	2.08	2.39	0.617
Race	1.28	1.31	0.949
Hispanic Origin	11.20	6.86	0.001

Table 8: Demographic Comparison of BCFE and Other Mail Return Populations in South Carolina

	Be Counted	Other Mail	
	%	%	p value
Sex	N=809	N=392,584	0.333
Male	47.96	46.26	
Female	52.04	53.74	
Age	N=808	N=387,830	0.006
Under 5	6.31	5.10	
5 to 14	15.47	13.48	
15 to 24	10.52	12.10	
25 to 44	28.47	29.66	
45 to 64	21.78	25.33	
65 and over	17.45	14.32	
Race	N=812	N=392,107	0.001
White	51.48	64.93	
Black or African American	46.18	32.70	
American Indian and Alaskan Native	0.62	0.38	
Asian	0.00	0.84	
Native Hawaiian and Other Pacific Islander	0.12	0.08	
Some other race	0.86	0.52	
Two or more races	0.74	0.55	
Hispanic Origin	N=751	N=370,068	0.124
Yes	1.07	1.82	
No	98.93	99.18	

Table 9: Sacramento Item Nonresponse Rates for 100% Person Data Items

	Be Counted	Other Mail	
	N=836	N=227,825	p value
Sex	3.83	1.47	0.001
Age	11.00	2.41	0.001
Race	16.99	9.02	0.001
Hispanic Origin	18.06	5.61	0.001

OUTMOVER TRACING FOR THE CENSUS 2000 DRESS REHEARSAL

David A. Raglin, Susanne L. Bean, United States Bureau of the Census
David Raglin; Census Bureau; Planning, Research and Evaluation Division; Washington, DC 20233

Key Words: Census coverage; Integrated Coverage Measurement (ICM); data collection; nonsampling error; measurement error; person matching

1. BACKGROUND

This evaluation provided information to help us determine if outmover tracing needs to be done as part of the Accuracy and Coverage Evaluation (A.C.E.) in Census 2000. Based on the results described here, the decision was made not to conduct outmover tracing in Census 2000. To aid in that determination, this evaluation answered the following questions:

- How many cases did we try to trace and what were the results?
- For households where a traced interview was obtained, how do the proxy and traced data compare?
- What is the person match rate to the census for the proxy data compared to the traced data?
- How are the estimates affected by replacing the outmovers provided by the proxies with the people provided by tracing outmovers?

The Census 2000 Dress Rehearsal was conducted in Columbia, South Carolina and eleven surrounding counties; Menominee County, Wisconsin; and Sacramento, California. Integrated Coverage Measurement (ICM) was the survey that followed the census and was designed to be a quality check survey on the census operations and to adjust census numbers. ICM was conducted independently of the census and collected an independent roster of residents as of census day and the ICM interview day.

2. METHODOLOGY

2.1 Definition of Movers

In the Census 2000 Dress Rehearsal, census day was April 18, 1998, and ICM data were collected via the computer-assisted personal interview (CAPI) Person

Interview from May to September, 1998. The problem is that people may have moved during that time. The people who have moved out of the housing unit after census day are called "outmovers".

To contrast, the people who have moved into the housing unit after census day are called "inmovers", and people who did not move between census day and the ICM Person Interview date are "nonmovers".

2.2 Effect of Movers on Estimation

The Census Bureau used Dual System Estimation (DSE) methodology to get adjusted population estimates based on census and ICM data, using information about both inmovers and outmovers collected during the ICM Person Interview. The DSE calculations are made within poststrata¹.

Below is the part of the DSE formula that is affected by movers: the term that measures the proportion of ICM people that matched census people:

$$\frac{M}{P} = \frac{M_{NM} + \frac{M_{OM}}{P_{OM}} \times (P_{IM} \times R P_{OM})}{P_{NM} + (P_{IM} \times R P_{OM})}$$

where:

- M = weighted estimate of people found in ICM who were matched to census people
 P = weighted estimate of people found in ICM
 M_{NM} = weighted estimate of the number of nonmovers found in ICM who match census persons
 M_{OM} = weighted estimate of the number of outmovers found in ICM who match census persons
 P_{OM} = weighted estimate of the number of outmovers found in ICM

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau Publications. This report is released to inform interested parties of research and to encourage discussion.

¹ The Dress Rehearsal poststratification variables were tenure (owner, renter), race/ethnicity (non-Hispanic White/Other, non-Hispanic Black, non-Hispanic American Indian/Alaska Native, non-Hispanic Native Hawaiian/Pacific Islander, non-Hispanic Asian, and Hispanic), and age/sex (0-17, 18-29 male, 18-29 female, 30-49 male, 30-49 female, 50+ male, and 50+ female).

- P_{DM} = weighted estimate of the number of in-movers found in ICM
- RP_{OM} = weighted estimate of the proportion of people enumerated as out-movers in ICM that were determined to be residents of the cluster on census day
- P_{NM} = weighted estimate of the number of non-movers found in ICM

Why do we use both in-movers and out-movers in the DSE equation? It is relatively easy to determine the number of in-movers from the ICM Person Interview, since we often are talking directly to them. Therefore, the number of movers within the poststrata is estimated based on the number of in-movers.

On the other hand, out-movers are people who lived in the housing unit on census day, so the out-movers are used to estimate the percentage of census and ICM matches.

The DSE methodology helps determine the focus of this evaluation. Out-movers are used in the DSE only to determine the match rate of movers to census people. Census and ICM people are matched using a clerical system, where people do not have to match on every data item to be matched.

2.3 Whole Household Out-movers

A whole household out-mover is a household where everyone moved out between census day and the ICM interview day. None of the people who lived there on census day are there when we go to do the ICM interview.

Why do we care specifically about whole household out-movers? Information about out-movers is needed so that they can be matched back to the data provided on the census form. If some of the residents had moved out but some were still there when the ICM Person Interview was conducted, the interviewer collected the information about the out-movers from the non-movers.

However, if everyone who lived there on census day has moved—a whole household out-mover situation—it is not quite as simple.

There are two options to collect information about whole household out-movers. One option is to find a knowledgeable proxy respondent to provide detailed information about the out-movers. Proxy information could be obtained from the in-movers or from neighbors or apartment managers who may have known the

out-movers. This proxy data was used in producing the official estimates for the Census 2000 Dress Rehearsal in Sacramento and Menominee and in measuring the undercount in South Carolina.

Another option is to attempt to trace the whole household out-movers to their new address and interview them about the household on census day. The advantages of this option are clear: theoretically, if the census day resident can be traced, they should know more about the census day household than a proxy would.

On the other hand, tracing out-movers can be difficult, time-consuming, and resource-intensive. In addition, proxies can often provide some of the needed information about the out-movers: name, age, sex, race, ethnicity, and relationship to the first person in the household so the clerical matchers can decide a match occurred. Out-movers are used in the estimation process to obtain the match rate to the census.

This evaluation is designed to determine if tracing whole household out-movers is worth doing by comparing the proxy data (which was used in the official dress rehearsal estimates) and traced data (collected especially for this evaluation).

2.4 Operation of Out-mover Tracing

There were two steps to out-mover tracing. First was an operation at the Census Bureau's National Processing Center (NPC) in Jeffersonville, IN. Researchers attempted to obtain the full name, telephone number, and new address of the out-movers (if the proxy had not already provided that information), utilizing all available resources, such as commercial databases, phonedisc and Directory Assistance, to find the necessary information.

Once the researchers got a valid phone number for the movers, they attempted to conduct a CATI interview. The CATI interview asked for the names and characteristics of everyone who lived at the sample address on census day and whether any of the persons had any alternate addresses on census day.

The second step was used if a case had not been found after fourteen days in the CATI unit. It got sent to the appropriate Field Division Regional Office and an interviewer tried to trace the movers to their new address by any means possible, such as knocking on doors and going to the post office. When interviewers successfully traced a mover, they conducted an

interview using a paper-and-pencil interview (PAPI) version of the CATI instrument.

3. LIMITATIONS

3.1 Operational Problems

Theoretically in outmover tracing, there never should have been a situation where we found that people never moved. In practice, it can be expected to happen a few times due to measurement error, but not often. However, it happened 15.4 percent of the time in Sacramento and 12.5 percent of the time in South Carolina for completed or resolved traced households. Unfortunately, if in the traced interview the person said they never moved from the Person Interview address, we did not follow up to resolve the discrepancy. An investigation spurred by this finding turned up specific problems that contributed to that and other problems with regard to outmover tracing. This makes the assumption that the Person Interview correctly identified whole household outmovers questionable.

It is important to note that these problems have been corrected for Census 2000.

3.2 General Limitations

The fact the dress rehearsal included only three sites in the country is a limitation in the ability to judge the PAPI operation. In Census 2000, there will be many outmovers in all parts of the country, and during a PAPI tracing operation, the PAPI forms will have to physically be sent around the country as more information is gathered about the outmover. Logistically, that is a problem. In dress rehearsal, we did not send PAPI cases out of the three sites. If a person, say, moved from Sacramento to Detroit but CATI could not find them in Detroit, when the case went back into the field, they only tried to trace them from Sacramento. In Census 2000, the case would be sent to Detroit and traced there.

However, the logistics of moving PAPI cases around the country during the census, which was not an issue during the dress rehearsal, would be very difficult. This issue has been legitimately raised as a reason not to conduct a PAPI operation in a census using outmovers.

In addition, there would be less time to trace outmovers via PAPI in Census 2000 than there was in the Census 2000 Dress Rehearsal. In dress rehearsal, tracing was conducted from June 19 to September 4. In Census

2000, PAPI tracing will take place from late June to late August. Also, because there are relatively few PAPI cases compared to the workload for other A.C.E. operations like Person Interview and Person Followup, and the PAPI cases will be scattered throughout both A.C.E. and non-A.C.E. clusters, it will be difficult for an interviewer to become especially skilled in tracing outmovers.

In addition, although these sites were chosen to represent situations found throughout the country, results of this evaluation cannot be generalized to any area beyond the three sites.

4. RESULTS

4.1 Amount of Outmover Tracing

How many households were there to trace? How well did we trace them? Table 4.1 answers that question:

Table 4.1: Results of Outmover Tracing

	Sacramento	S. Carolina
ICM Housing Units	16,419	17,677
Whole HH Outmov.	918 (6%)	927 (5%)
Traced Households	380 (42%)	482 (52%)

Data for Menominee are not provided since there were only 10 whole household outmovers for the site. The households who reported they never moved are not included in the "Traced Households" line because a large percentage of them should not have gone to outmover tracing, as explained in section 3.1 of the limitations. They made up 7.5 percent of the Sacramento cases and 7.4 percent of the South Carolina cases that went to outmover tracing.

The indication from these results is that about five to six percent of households in Sacramento and South Carolina were whole household outmovers. However, the problems in the limitations section indicate that some of the cases that went to outmover tracing should not have, so the percentage of true whole household outmovers is probably a little less than five percent in those sites.

Among the total tracing workload, 26 percent were traced via the CATI system and 16 percent via PAPI in Sacramento. In South Carolina, 40 percent of the total workload was traced via CATI versus 12 percent by PAPI.

CATI tracing in Sacramento might have been hampered by the large percentage of unlisted phone numbers there.² In the debriefing of the CATI interviewers, the interviewers mentioned they could not get addresses for people with unlisted telephone numbers from directory assistance, making it harder to trace those people (Ehni, 1998). However, the number of cases traced in Sacramento indicates that they nonetheless were traceable, just not via a telephone operation.

This indicates that the effectiveness of CATI tracing could vary for different parts of the country, an important finding for planning a census using outmover tracing.

4.2 Comparing Proxy and Traced People

In households we traced, we compared the people found in the ICM proxy interview with the people found in the ICM traced interview. The goal was to determine if we were getting the same people and households in the traced interview as in the proxy one. After all, if we trace and the household list of people contains the same people, tracing is not particularly useful. In 1996, a similar matching operation found that in Chicago, 21.9 percent of the time when we traced a household, we actually got an entirely different list of people.

This matching took into consideration the use of outmover tracing with the DSE methodology in mind. DSE uses outmovers only to compute the match rate between the census and the ICM. Therefore, Census Bureau clerical person matching rules were used in deciding if a proxy and traced person were the same person. A person experienced with the clerical person matching matched the proxy and traced people. Matches were attempted only for data-defined people with a valid name.³

Table 4.2 shows the results of the proxy versus traced matching in households where the household was

traced and an interview was obtained that included the listing of people.

Table 4.2: Comparing Proxy and Traced People

Person Mentioned In	Sacramento	
	Proxy	Traced
Both proxy & traced ints.	431	431
Only in proxy interview	31	---
Only in traced interview	---	233

Person Mentioned In	South Carolina	
	Proxy	Traced
Both proxy & traced ints	528	528
Only in proxy interview	46	---
Only in traced interview	---	306

If the person was mentioned in the proxy interview, they were almost always found in the traced one too (93 percent in Sacramento ($431 / (431 + 31)$) and 92 percent in South Carolina). If we assume that the traced interview is better than the proxy interview, since it was supposed to be with a resident, a large percentage of the people that proxies name are really residents.

Table 4.2 also indicates the traced interview found many additional people that the proxy interview did not have. Since the traced interview was supposed to have been done with a resident of the outmover household, this is not a surprise.

In which households were these new traced people found? One theory was that the new people found in the traced interview were in households where we did not collect any proxy people. However, analysis showed that was not true. In Sacramento, 133 of the 233 additional people (57.1 percent) from the traced interview were in households where we got people in the proxy interview who were not in-movers; for South Carolina, it was 144 of 306 (47.1 percent). Therefore, even if we get people in the proxy interview, we get even more outmovers in the traced interview.

The proxy interview seems to yield legitimate people but an incomplete list of the household members (at least according to the traced interview).

² Approximately 71.6 percent of the households with telephones in the Sacramento PMSA have unlisted numbers, the largest percentage of any of the biggest 100 metropolitan area in the country (Survey Sampling, Inc. 1999).

³ To be data-defined, we must have a valid name and one other characteristic. The name must have at least three characters in the first and last name together. The characteristics include relationship, sex, race, Hispanic origin, and age or year of birth or month and day of birth.

4.3 Matching to Census People

Remember that under the DSE methodology used, outmovers are included for their person matching rate to the census. The number of movers is estimated by the inmovers. Therefore, the person match rate is a very important indication of the difference in data quality between the proxy and traced data.

Table 4.3 shows the match rate for the data using proxies for outmovers compared to the data using the traced data for outmovers when traced data was available. That means that if a traced interview was conducted and people were collected, the traced interview replaces the proxy interview. If the traced interview indicated the housing unit was vacant or did not exist as a housing unit on census day, the production people were removed.

Table 4.3: Match Rates to Census People

ICM Persons	Sacramento	
	Proxy	Traced
In traced households	65.1	65.6
In all of ICM	78.2	78.2

ICM Persons	South Carolina	
	Proxy	Traced
In traced households	79.0	75.5
In all of ICM	78.5	78.5

Look at the first line of the table in Sacramento, the people in households where we collected person data during tracing. Notice that the match rates are almost the same for the proxy people as for the traced people. In the traced interview, we find many more people, but their match rate is similar to the match rate for the people we already had. In South Carolina, we actually had a nominally lower match rate for the people collected in the traced interview.⁴

The last line of the table in each site shows that to one decimal place, the match rate was the same for the dataset that used proxy people versus the traced people. This indicates that even though we find quite a few

⁴ The match rates for proxy versus traced people are similar for the poststrata marginal variables, too.

new people in the traced interview, they do not match at a particularly high rate to census people. That is due to the fact that only about two percent of households were whole household outmovers that we were able to trace as well as the similarity of the match rates between proxy and traced people.

4.4 DSEs Using Proxy and Traced People

This is really the most important section of this paper. We can say the match codes do not seem to differ between the proxy and traced people, but if there are significant differences in the DSEs, we can say the differences in the match rates were actually significantly large.

Table 4.4 shows the DSEs calculated from the production data (excluding people in groups quarters and in the service-based enumeration), the DSEs calculated from the data using the traced outmovers in place of the proxy outmovers in households we were able to trace, the differences, and whether or not those differences are significantly different than zero.

Table 4.4: DSEs Using Proxy and Traced People⁵

	Sacramento	S. Carolina
Est.w/ proxy people	395,005	693,724
Est w/ traced people	395,025	693,579
Difference (st error)	20 (279)	-145 (522)
Significant ($\alpha = .10$)	No	No

We also did this comparison for each poststrata marginal variable using the Dunn method of controlling for multiple comparisons.⁶

There were not significant differences in the DSEs calculated using proxy and traced outmover people in either site for any of the poststrata marginal variables, with $\alpha = .10$. In fact, the p-values are not close to being significant most of the time. Outmover tracing

⁵ The numbers here are the estimates for the whole site minus people collected from group quarters and the service-based enumeration operation.

⁶ In the Dunn method, the alpha level was divided by the number of comparisons to be made: one for the total, two for tenure, seven for race/ethnicity, and six for age/sex, to come up with the significance level used in the tests. See Toothaker (1993).

provided a nominal increase of 20 people in Sacramento (0.005 percent of the production estimate) and a nominal loss of 145 people in South Carolina (0.021 percent of the production estimate).

There is no reason to believe that with the current DSE methodology, the lack of tracing caused any significant difference in the production estimates provided in the Census 2000 Dress Rehearsal.

5. ANALYSIS AND CONCLUSIONS

The big question is whether or not the outmover tracing operation should be included as part of the A.C.E. in Census 2000. Of course, this evaluation is based on data from two sites -- Sacramento and the parts of South Carolina that were included in the Census 2000 Dress Rehearsal. It could be possible that outmover tracing could have a significant effect on estimates in parts of the country that differ from these two sites.

It is understandable why outmover tracing might not have a significant effect on the estimates. Remember that outmovers are used to produce an estimate of the match rate between the census and the ICM. Look back to Table 4.2, the table that showed the results of matching the proxy and traced people to each other. We found that almost all of the time if a person was mentioned by a proxy and they knew enough about the person that we could consider them data-defined, the person was later mentioned in the traced interview.

To contrast, the people mentioned by the proxy but not the traced respondent might logically be assumed to not be residents, and in fact they did not match well to the census: 70 percent did not match in Sacramento and 43 percent did not match in South Carolina. However, there are so few of them that their effect on the match rate is relatively small.

Therefore, for the people in the traced interview to have a much higher match rate than the proxy interview people, the new people mentioned in the traced interview but not the proxy one would have to match to the census at a higher rate than the people mentioned in both interviews. There is no reason to think that to be true, and in fact it isn't.

From the people in the traced interview, 32 percent of the people also mentioned in the proxy interview did not match anyone in the census in Sacramento, while 39 percent of the people mentioned in the traced interview but not mentioned in the proxy interview did not match census people. In South Carolina, the

figures were 20 percent versus 30 percent.

While it would be nice to have all of the people in the household (as tracing would help us do), we really don't care about the number, just the match rate. The proxies seem to be giving us good enough data for matching purposes.

We therefore recommended that outmover tracing not be conducted as part of the Census 2000 Accuracy and Coverage Evaluation. Based on this analysis, this recommendation was accepted, and there will be no tracing of outmovers in Census 2000.

References

Ehni, Sandy (1998), "Debriefing Summary for the Census 2000 Dress Rehearsal of the Integrated Coverage Measurement Outmover Tracing Operation", Internal Census Bureau memorandum.

Survey Sampling, Inc (1999), "Unlisted Rates", [http://www.ssisamples.com/ssi.x2o\\$ssi_gen.product?id=71](http://www.ssisamples.com/ssi.x2o$ssi_gen.product?id=71).

Toothaker, Larry (1993), "Multiple Comparison Procedures" (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-089), Sage University Press, Newbury Park, CA.

Census 2000 Dress Rehearsal
Zakiya T. Sackor
US Bureau of the Census, Rm BH119-2, Washington, DC 20233

Key words: Long form sample loss, alternate data collection forms

Introduction

The Census 2000 Dress Rehearsal is the Census 2000 testing program that was conducted in Columbia, South Carolina and eleven surrounding counties; Menominee County, Wisconsin; and Sacramento, California. Only two of the Dress Rehearsal sites will be examined in this analysis, Sacramento and South Carolina.

Two forms were used to collect Dress Rehearsal data--the short and long forms. The short form contains basic demographic questions such as sex, race/ethnicity, relationship, length of time lived at their residence, and the number of people living at the residence. The questions on the short form are known as the 100% items. In addition to the short form items, the long form contains questions such as marital status, highest level of education completed, income, citizenship, employment, and housing items. The household's a priori form assignment is determined by a sampling scheme, that assigns on average 1-in-6 households a long form, and all others households a short form.

This paper discusses two factors that contributed to the loss of long form sample data in Dress Rehearsal: 1) the use of alternate data collection forms and 2) total nonresponse to all long form items. For the Dress Rehearsal, respondents were allowed to respond through two alternate data collection methods: the Be Counted Form (BCF) and interviews taken through Telephone Questionnaire Assistance (TQA). There are also cases in which a return was received without any long form data. Each factor will be discussed independently.

Alternate Data Collection Forms

Background

One of the goals of Census 2000 is to maximize response, both through the use of respondent friendly forms and through more accessible alternate data collection forms. There are two alternate data collection forms--Be Counted Forms (BCF) and

interviews taken through Telephone Questionnaire Assistance (TQA).

There are several possible situations which could lead respondents to use Be Counted Forms (BCF) or the TQA data collection instrument. First, the BCF was designed to provide an opportunity for people to be counted who think that they were not included on any other census form. For example, if a boarder in a single family household was not listed as part of that household on the mail form, he/she could pick up a BCF, complete and return it. Though not the original intention of TQA, it can accommodate this situation as well.

Second, the TQA interview instrument was developed to help people respond who for any reason were having difficulty completing their mail form by giving them the opportunity to provide the census information over the phone with the assistance of a trained interviewer. This includes people who have more than one place to live and are not sure at which residence to count themselves, as well as people who have misplaced their mail forms, never received a form, or received a form with an incorrect address. People who have misplaced their form, never received a form, or received a form with the incorrect address can also request another form be mailed to them. For the purpose of this analysis, when the caller cannot provide their census identification number in either an interview or replacement mailing situation, it will be considered an alternate data collection. Without an identification number it is not possible to identify the form type assigned to the caller's address, so a form type is assigned within the TQA instrument (i.e., short to long forms). Thus, it is possible in either the TQA interview or TQA replacement mail situation to collect short form data from housing units that were intended to be long form recipients, and vice versa.

The objective of this evaluation is two-fold, 1) to determine the percentage of households and people who are counted by a short form rather than a long form as intended; 2) to determine whether there are significant statistical differences between those who

respond by long form as intended and those who do not.

Methodology

The first step of the analysis compared percentages of final selected returns¹ that were a priori long form but were counted by a short form at the household level. The second step included a test of differences for age and race/ethnicity.

Results

There were 19,469 households in Sacramento that were a priori long form, of these:

- 19,196 (98.6%) were counted as assigned
- 62 (0.4%) were counted by an alternative data collection method
- 211 (1.0%) were counted by a short form mail or enumerator return

There were 37,223 households in South Carolina that were a priori long form, of these:

- 36,847 (99.0%) were counted as assigned
- 29 (0.1%) were counted by an alternative data collection method
- 347 (0.9%) were counted by a short form mail or enumerator return

The last bullet in each of the sites show that there are instances in which a household can return either a short mail or enumerator form, such as the case when 1) multiple returns are received from a household, and a short form is selected over a long form or vice versa, 2) through the mislabeling of forms during Nonresponse Followup or 3) households added during Update Leave that were given a short form instead of a long form.

¹ More than one return may be returned and checked-in for an address. However, the Census Bureau employs an algorithm based on several factors that selects the final return(s) for an address. This analysis is based only on the selected returns.

Test of Difference

Tables 1 - 2 include a test of difference for age and race/ethnicity of a priori long form persons that were counted as assigned and those that were not. Each column includes all form types- mail returns, update leave (South Carolina only), enumerator returns, and alternate data collection forms. Only significant statistics are shown in the table.

TABLE 1 -DEMOGRAPHICS OF PERSONS WHO WERE COUNTED AS ASSIGNED VERSUS NOT AS ASSIGNED

SACRAMENTO		
Demographic Characteristics	Counted as Assigned	Not Counted as Assigned
	N= 43,796	N=848
Age		
< 6	8.32%	11.79%
6 - 17	17.37	20.52
18 - 29	15.33	19.22
30 - 45	25.13	25.47
46 - 65	20.50	14.62
> 65	13.35	8.37
Race/Ethnicity		
White, alone	54.94%	39.29%
Black, alone	18.09	19.75
American Indian/ Alaska Native alone	2.78	3.07
Asian	14.41	21.46
Two or more races	4.96	6.49
Hispanic or Latino (of any race)	19.03	26.18

$p < .001$

Those in the younger categories, below the age of 30, were more likely to not be counted as assigned.

Races other than White and Hispanics appear less likely to be counted as assigned compared to Whites.

TABLE 2 - DEMOGRAPHICS OF PERSONS WHO WERE
COUNTED AS ASSIGNED VERSUS NOT AS ASSIGNED
SOUTH CAROLINA

Demographic Characteristics	Counted as Assigned N= 86,582	Not Counted as Assigned N=963
Age		
< 6	7.88%	9.87%
6 - 17	17.01	20.52
18 - 29	15.46	19.22
30 - 45	24.74	25.03
46 - 65	22.76	21.50
> 65	12.14	7.48
Race/Ethnicity		
White, alone	57.92%	51.09%
Black, alone	37.79	45.59
Hispanic or Latino (of any race)	1.99	1.14

$p < .001$

- Those below the age of 30 appear more likely to not be counted as assigned.
- In several categories it appears that races other than White were less likely to be counted as assigned compared to Whites.

Conclusions

- Alternative data collection methods had virtually no effect on long form sample loss, about 0.4% in Sacramento and 0.0% in South Carolina.
- The majority of long form sample loss is due to form assignment problems with added units during update/leave or persons being counted by an enumerator short form.
- The distribution of long form persons not counted as assigned was greater for races other than White.
- The TQA operation for Census 2000 is currently being redesigned to collect only short form data. Although this new design has the potential to further increase long form sample loss due to alternate data collection forms, we do not anticipate this as a major concern since the Dress Rehearsal results indicate that a very few households were counted through a TQA long form interview.

Nonresponse to Long Form Items

Background

In 1990 the Census Bureau experienced substantial loss of long form data due to total nonresponse to the long form items. These forms were essentially converted to short forms. In 1990, there were two forms received during a Census-- mailout/mailback forms and enumerator forms. Enumerator forms were used during Nonresponse Followup (NRFU), an operation in which data was collected from households that did not originally respond to the Census.

In 1990 over 22% of the enumerator long forms had to be converted to short forms because they were missing all of the long form data. Only 1% of mail long forms were converted. The design of the 1990 enumerator long form may have been a contributing factor to the loss of long form sample data. The design was as follows: Short form person items, long form housing items, and finally long form person items. The data from 1990 indicates that enumerators were completing the short form person and housing items, but in many cases were not able to get any long form person data. Based on this, the analysis in 1990 examined long form sample loss with respect to long form person items only.

Methodology

In the Dress Rehearsal (DR) the long form was redesigned to begin with the short form person items followed by the long form person items, and long form housing items. Total nonresponse for the long form person items were calculated for all data defined persons within a household. If the entire household completed at most one long form item this was considered nonresponse to the long form. The methodology also includes a test of difference for age and race/ethnicity for those who responded to two or more person items versus those who did not. The test of differences was restricted to enumerator forms only.

Results

The column categories indicate the number of converted mail and enumerator returns from 1990, as well as the percentage of forms missing all long form person items for each of the Dress Rehearsal sites. The mail and enumerator categories for 1990 include the percentage of converted forms for the entire state (i.e.

California and South Carolina). The Dress Rehearsal data are only comprised of the city of Sacramento and 11 counties surrounding and including Columbia, South Carolina.

TABLE 3- PERCENTAGE OF MAIL AND ENUMERATOR FORMS MISSING ALL LONG FORM ITEMS FROM 1990 VS DRESS REHEARSAL

Site	Mail		Enumerator	
	1990 (state level)	DR	1990 (state level)	DR
Sacramento	1.1 (16,145)	.3 (36)	25.1 (368,399)	22.2 (1,194)
South Carolina	1.2 (2,401)	.4 (88)	22.8 (45,610)	13.2 (1,669)

The table above indicates that in Sacramento, there was about a 3% decrease in long form sample loss from enumerator forms in the DR when compared to 1990, while South Carolina observed approximately a 10% decrease in long form sample loss from 1990.

Test of Difference

Only significant statistics are shown in Tables 4-5.

TABLE 4 - DEMOGRAPHICS OF HOUSEHOLDERS MISSING ALL LONG FORM ITEMS VERSUS THOSE THAT DID NOT SACRAMENTO

Demographic Characteristics	All Long Form Items Missing	Did Not Have all Long Form Items Missing
	N= 1,194	N=4,162
Race/Ethnicity		
White, alone	54.94%	39.29%
Black, alone	18.09	19.75
Native Hawaiian/ Other Pacific Islander alone	9.05	11.27
Hispanic or Latino (of any race)	14.74	22.13

$p < .001$

- It appears that Whites are more likely to complete a most one long form item.

TABLE 5-DEMOGRAPHIC COMPARISONS OF HOUSEHOLDERS MISSING ALL LONG FORM ITEMS VERSUS THOSE THAT DID NOT SOUTH CAROLINA

Demographic Characteristics	All Long Form Items Missing	Did Not Have All Long Form Items Missing
	N= 1,669	N=10,983
Age		
< 6	0.12	0.02
6 - 17	0.48	0.15
18 - 29	20.91	18.02
30 - 45	42.24	39.83
46 - 65	26.36	29.41
> 65	9.89	12.56
Race/Ethnicity		
White, alone	52.61%	48.65%
Black, alone	42.72	47.42
Hispanic or Latino (of any race)	2.70	2.050

$p < .001$

- Those older than 45 are more likely to respond to two or more long form person items.
- It appears that Whites are more likely to complete at most one long form items.

Conclusions

Mail back forms had little or no effect on long form sample loss. We expect nonresponse to the mailed back forms to be low because respondents who take time to return the questionnaire have probably completed it. Enumerator forms, on the other hand, significantly contributed to long form sample loss. Compared to 1990, there was a decrease in total long form nonresponse for enumerator forms in South Carolina; this may be a result of the redesign of the long form. After the 100% items which are asked for each person in the household, the redesign allowed respondents to answer person items for each person within the household, instead of moving directly to the housing

items as in 1990. This smooth transition from the 100% items to the person items may be the reason for the decrease in total long form nonresponse for enumerator forms in South Carolina. However, in Sacramento the nonresponse rate remained relatively stable; this may be a result of the urbanicity of Sacramento. Asians and Hispanics account for about 36% (144,721) of the population in Sacramento, but only about 3% (18,617) of the population in South Carolina. English may be a second language to many residents of urban areas such as Sacramento. These individuals may not respond to the Census because they may have some level of distrust for the government or may simply be unaware of the Census, and its benefits to their communities. The Census Bureau is currently examining partnership and advertising programs specific to these communities, to increase the awareness and importance of the Census.

During NRFU, enumerators are to collect data from respondents that did not respond to the Census initially. The Census Bureau must equip enumerators with facts about the importance of the Census, along with the notion that overall only 1 in 6 households receive the long form, therefore it is imperative that quality data be collected from the selected households. Enumerators should also be given extensive persuasion tactics to obtain sufficient data from persons that are not eager to respond.

In South Carolina, significant statistical differences were found for those in older age groups that completed two or more long form person items and those who did not. Specifically, the data indicates that persons over 45 are likely to complete more long form person items than younger individuals. The data also suggest that statistical differences were found for the race/ethnicity distributions for those who completed two or more long form person items and those who did not. The data show that Whites are more likely to complete at most one item, where minority groups complete two or more items.

There are several reasons for nonresponse to the Census, from its seemingly intrusive nature to the design of the questionnaire. All avenues that contribute to nonresponse should be explored extensively for the success of the Census.

Woltman, H. (1990) STSD 1990 Decennial Census Memorandum

NOTE: This paper reports the research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau Publications. This report is released to inform interested parties of research and to encourage discussion.

References

Iterative Proportional Fitting in the Census 2000 Dress Rehearsal
Eric Schindler, Bureau of the Census, Washington DC 20233

KEYWORDS: Dual System Estimation,
Poststratification, Raking

Iterative proportional fitting, or raking, was employed in addition to the dual system estimation methodology to measure the undercoverage for the Census 2000 Dress Rehearsal conducted during 1998 in three sites. The raking procedure was used to adjust the initial phase estimates for poststrata defined by race/origin/age/sex/tenure to two sets of marginals defined by race/origin/age/sex and tenure estimated by taking the sums of direct dual system estimates for the same poststrata. This procedure was designed specifically to improve reliability and preserve the race/origin/age/sex cells required for congressional and state redistricting and to induce approximately the same coverage differences between owners and renters for each demographic group. This paper discusses the results of the procedure and of several alternative raking matrices with a view towards Census 2000.

I. Introduction

Two years before each decennial census, the Bureau of the Census executes a Dress Rehearsal, a full scale implementation of the census in several small sites. The dress rehearsal is designed as a final test for operations, forms, estimation, and data publication. In theory, after the dress rehearsal only minimal adjustments to correct serious shortcomings should be implemented. The political controversy surrounding Census 2000 has made it difficult for the Census 2000 Dress Rehearsal to follow this prescribed course.

Two of the three Dress Rehearsal sites, Sacramento, CA and Menominee County, WI, mostly an American Indian reservation, used the Integrated Coverage Measurement (ICM) design, which based on a recent Supreme Court ruling will not be used for reapportionment in 2000. The third site, Columbia, SC and several surrounding counties, was collected as a

Eric Schindler is a mathematical statistician in the Decennial Statistical Studies Division of the US Census Bureau. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. Research results and conclusions expressed are those of the authors and do not necessarily indicate concurrence by the Census Bureau. It is released to inform interested parties of current research and to encourage discussion.

traditional census followed by a Post Enumeration Survey (PES). While the operations and results in Sacramento and Menominee were much as expected, it appears that there was substantial undercoverage in the address listings in the mail delivery areas outside of Columbia, leading to higher than expected estimated undercount rates for owners.

For all sites, the coverage survey was collected in a stratified (by predominant race and home ownership) sample of blocks or block clusters averaging about 30 housing units. The sampled housing units were reinterviewed independently. The persons counted in the initial phase and coverage collection efforts were compared to determine which persons in the initial phase were correctly enumerated and which persons in the coverage survey could be matched to the initial collection.

Estimation of coverage for both designs was by a poststratified dual system estimator. Poststrata were defined by 6 race/Hispanic origin groups, 7 age/sex groups, and tenure, the owner/renter dichotomy. Raking, or iterative proportional fitting (first suggested for the Decennial Census in Schindler and Griffin, 1997), was implemented on a 42 by 2 matrix to help control the standard errors.

Section II discusses the results of the estimation and raking procedures at the site and poststratum level. Section III examines several alternative estimation poststratification and raking options. Section IV concludes the paper by discussing current plans for Census 2000.

II. The Census 2000 Dress Rehearsal

POSTSTRATIFICATION

Poststrata were defined using race, Hispanic origin, age, sex, and tenure. These variables, plus geographic identifiers had been used in the 1990 PES. The six race/Hispanic origin groups are:

1. Non-Hispanic White/Other
2. Non-Hispanic Black
3. Non-Hispanic American Indian/Alaska Native
4. Non-Hispanic Native Hawaiian/Pacific Islander
5. Non-Hispanic Asian
6. Hispanic (of any race)

Persons who marked more than one race box were assigned during estimation to the largest non-white race marked based on 1990 site level census counts. Race/origin groups which were less than 1% of the site total in 1990 were collapsed during estimation into the

largest non-white race based on the 1990 census counts. In Sacramento, American Indians and Native Hawaiians were combined with Hispanics. In Columbia, all minority races were combined with Blacks. In Menominee, Blacks, Native Hawaiians, and Asians were combined with American Indians.

The seven age/sex groups were:

0-17	18-29 Male	18-29 Female
	30-49 Male	30-49 Female
	50+ Male	50+ Female

In the Menominee site some of the age/sex categories were combined for White/Others and for Hispanics.

DUAL SYSTEM ESTIMATION

The site level dual system estimate for the three Census 2000 Dress Rehearsal Sites was defined by:

$$D\hat{S}E_i = C_i \times \frac{\hat{D}_i}{\hat{C}_i} \times \frac{C\hat{E}_i}{\hat{E}_i} \times \frac{\hat{P}_i}{\hat{M}_i}$$

where: C_i is the initial count in poststratum i.

\hat{C}_i is the initial count in poststratum i as estimated from the initial count in the coverage sample areas

\hat{D}_i is the estimated number of initial count persons for whom at least some data was directly collected - only data defined persons are included in the E-Sample

\hat{E}_i is the estimated number of persons in the enumeration sample in poststratum i. This is approximately equal to \hat{D}_i

$C\hat{E}_i$ is the estimated number of E-Sample persons in poststratum i who are determined to have been correctly enumerated in the initial collected effort.

\hat{P}_i is the estimated number of persons in the coverage survey in poststratum i.

\hat{M}_i is the estimated number of P-Sample persons in poststratum i who can be matched to a person in the initial collection effort.

Because of the special treatment required for persons who move between Census Day and the time of the coverage interview, the last term of the equation is actually somewhat more complicated:

$$\frac{\hat{P}_i}{\hat{M}_i} = \frac{\hat{P}_{i,nonmover} + \hat{P}_{i,inmover} \times \frac{\hat{P}_{i,outmover}}{\hat{P}'_{i,outmover}}}{\hat{M}_{i,nonmover} + \hat{P}_{i,inmover} \times \frac{\hat{M}_{i,outmover}}{\hat{P}_{i,outmover}} \times \frac{\hat{P}_{i,outmover}}{\hat{P}'_{i,outmover}}}$$

where nonmover, inmover, and outmover are obviously defined and $\hat{P}'_{i,outmover}$ is the estimated number of persons originally enumerated in the P-Sample including some

who are later determined to have not been residents. This treatment of movers, known as DSE C, uses the best available data to estimate the number of movers (from the in-movers) and the best available data to estimate the residence and match probabilities (from the out-movers, collected by proxy).

RAKING

A simple raking or iterative proportional fitting procedure was executed to help control the variances. The raking matrix was defined by the collapsed race/Hispanic origin by age/sex categories in the first dimension and tenure in the second dimension. The marginal controls were calculated by adding the dual system estimates for the interior cells. The initial phase estimates were then raked to the marginal controls.

The variance estimates in this paper were estimated by a simple Jackknife procedure which gives similar results to the stratified Jackknife used in the official estimates. Bias² at the poststratum level for the raked alternatives is estimated by: $(DSE-Rake)^2 - VAR_{DSE-Rake}$ and the estimates are often negative; MSE is estimated by $VAR_{Rake} + Bias^2$ and the estimates are sometimes negative.

DRESS REHEARSAL SUMMARY RESULTS

Table 1 shows some summary results, excluding persons in group quarters, of the initial phase, dual system estimates and raked estimates for the three Dress Rehearsal sites. Past experience has shown that coverage in test censuses is generally worse than in the actual census.

Table 1: Summary Results: Dress Rehearsal Estimation

Sacramento	Initial	DSE	%UC	rakeDSE	%UC
Total	369434	395005	6.47%	395005	6.47%
owner	188202	194398	3.19%	194398	3.19%
renter	181232	200608	9.66%	200608	9.66%
White/Other	160620	168555	4.71%	168555	4.71%
owner	91545	94020	2.63%	93703	2.30%
renter	69075	74535	7.32%	74851	7.72%
Black	59005	64647	8.73%	64647	8.73%
owner	22355	23141	3.40%	23350	4.26%
renter	36650	41506	11.70%	41297	11.25%
Asian	58890	62643	5.99%	62643	5.99%
owner	33057	34153	3.21%	34151	3.20%
renter	25833	28490	9.32%	28492	9.33%
Hispanic +	90919	99161	8.31%	99161	8.31%
owner	41245	43083	4.27%	43193	4.51%
renter	49674	56078	11.42%	55967	11.24%

Menominee	Initial	DSE	%UC	rakeDSE	%UC
Total	4550	4694	3.06%	4694	3.06%
owner	2937	3026	2.93%	3026	2.93%
renter	1613	1668	3.30%	1668	3.30%
Amer Ind +	3859	4024	4.10%	4024	4.10%
owner	2327	2485	6.34%	2450	5.02%
renter	1532	1540	0.49%	1574	2.68%
Columbia	Initial	DSE	%UC	rakeDSE	%UC
Total	628616	693724	9.39%	693724	9.39%
owner	452310	507865	10.94%	507865	10.94%
renter	176306	185858	5.14%	185859	5.14%
White/Other	359854	384073	6.31%	384073	6.31%
owner	286891	308384	6.97%	310670	7.65%
renter	72963	75688	3.60%	73403	0.60%
Black +	268762	309651	13.20%	309651	13.20%
owner	165419	199481	17.08%	197195	16.11%
renter	103343	110170	6.20%	112455	8.10%
+ : In Sacramento Native Hawaiians and American Indians were collapsed with the Hispanic population. In Menominee, Blacks, Asians, and Native Hawaiians were collapsed with the American Indian population. In Columbia all groups except non-Hispanic Whites were collapsed with the Black population. Rounding may cause slight differences between the published results and those presented here.					

The results for Sacramento were consistent with those observed in the 1990 PES. The estimated undercount rates were higher for renters than for owners and higher for minorities than for Whites. The total undercount (6.47%) was higher than the 1990 PES, but this is consistent with experience from the 1990 Dress Rehearsal. The raking procedure widened the differential between White owners and renters and narrowed it between Black or Hispanic owners and renters. The standard errors for the raked poststrata average about 80% of the preraking values. The unweighted mean square errors of the 56 postcollapsing poststrata after raking average only 22% of the unraked poststratum variances, indication significant bias reduction.

Although similar 17% and 46% reductions in the average poststratum level standard errors and mean square errors, respectively, occurred, the estimation in the Columbia site did not proceed nearly as smoothly as in Sacramento. The overall undercount rate was about 3% higher. The undercount rate for minorities (almost all Black) was higher than that for Whites, but the undercount rate for owners was higher than that for renters for both race groups. This was especially so for the poststrata for owners under 50 years old. The poststrata for white older persons showed the expected higher undercount rates before raking for renters than for owners. For minorities, the owners over 50 years old had

higher undercounts than the over 50 renters but not by as much for the under 50 owners.

Table 2: Undercount Rates by Tenure in Columbia

Race	Age	Owner	Renter
Non-Hispanic White	<50	9.22%	2.57%
	≥50	3.25%	6.43%
All Others	<50	18.78%	5.88%
	≥50	11.74%	5.54%

Investigation showed that the estimates for just the city of Columbia were consistent with those for Sacramento. Also, the estimates for the most rural areas of the site where the Census Bureau enumerators updated the address list while delivering forms were also acceptable. The bulk of the problem occurred for the 60% of the site population who lived in mailout/mailback areas outside of Columbia where the addresses were provided by the Postal Service. It appears that many housing units near the edges of these mailout/mailback areas were not reported. These units are mostly owner occupied. This problem will be addressed in Census 2000 by redesigning the creation of the Master Address File, a major component of which will be a block canvassing operation searching for additional housing units nationwide.

Seven block clusters (out of 665 total block clusters) with varying collection problems, all in the mailout/mailback areas outside of Columbia or the update/leave areas, had a disproportionate impact on the estimates. Table 3 shows the Columbia estimates omitting these seven block clusters. The owner/renter reversal has been greatly reduced, from about 5.8% to 0.7% at the site level and from 8.1% to about 2.8% at the poststratum level.

Table 3: Columbia Results Omitting Outliers

Columbia	Initial	DSE	%UC	RakeDSE	%UC
Total	628616	689593	8.84%	689593	8.84%
owner	452310	497283	9.04%	497283	9.04%
renter	176306	192310	8.32%	192311	8.32%
White	359854	383402	6.14%	383402	6.14%
owner	286891	306135	6.29%	306585	6.42%
renter	72963	77266	5.57%	76817	5.02%
Black	268762	306192	12.22%	306192	12.22%
owner	165419	191147	13.46%	190698	13.26%
renter	103343	115044	10.17%	115494	10.52%

The reversal of undercoverage rates for owners and

renters from the expected direction in Menominee has not been thoroughly investigated.

III. Design Alternatives

Raking, as applied in the Census 2000 Dress Rehearsal, approximately imposed a uniform difference in estimated undercount rates on all pairs of owner/renter poststrata. This difference was about 5.5% in Sacramento and -8.75% in Columbia. It is possible to lessen this consistent effect by defining more marginal cells in the second dimension of the raking matrix. Recall that the first dimension was defined by the six race/Hispanic origin categories crossed by the seven age/sex categories which were then collapsed to eliminate small cells. The second dimension was defined by the two tenure cells only. Two types of additional variables can be defined in the second dimension. The first is designed to capture interactions between race/origin or age/sex and tenure by using reduced versions of the variables in the first dimension. The second type of variable, "new" variables, would allow the formation of more homogeneous poststrata.

A. Interactions

A race/age/tenure interaction was observed in Columbia which was masked by the raking. Therefore, two additional dichotomous variables were defined: over/under 50 and nonminority/minority. These additional variables can be included in the second dimension of the raking matrix, yielding eight marginal cells instead of two. Since these variables are already included in the first dimension of the raking matrix in an expanded form there is no change to the before raking estimates at summary levels at the race/origin by age/sex level. The addition of these variables to the second dimension of the raking matrix retains most of the gains of raking for standard errors and reduces the changes made from the before raking poststratum estimates.

The estimates for Black children in Sacramento are typical, with lower MSEs and almost 100 fewer persons moved from renter to owner when the interactions are included. With the interactions in the raking procedure, the estimates for Black children are less biased (in fact, the estimate of BIAS² is negative), have lower mean square errors, and preserve almost all of the reduction in standard error of the simple raking procedure.

Table 4A: Estimates for Black Children in Sacramento

Estimate	Owner	Renter
Initial Estimate	6477	14855
DSE before Raking (se) Undercount Rate	6624 (265) 2.22%	17203 (747) 13.65%
Rake by Tenure (se) Undercount Rate RMSE / BIAS Difference from DSE	6942 (204) 6.70% 259 / 161 +318	16885 (605) 12.02% 626 / 161 -318
Rake by Tenure, Minority Status, and +/- 50 (se) Undercount Rate RMSE / BIAS Difference from DSE	6857 (209) 5.54% 190 / n/a +233	16971(628) 12.47% 621 / n/a -233

The real advantage of including the interactions occurs in Columbia where the results were not as expected. White owners had worse coverage than White renters, but not White owners over 50. Raking by just tenure forced the older population to follow the overall pattern. For White males over 50, 781 persons, 1.6% of the before raking DSE, were moved from the unraked renter poststratum to the raked owner poststratum and error estimates increased substantially with large biases. Raking with race and age as well as tenure in the second marginal includes the interactions, stops the imposition of average coverage factors, significantly reduces the bias, reduces the MSEs especially for renters, and reduces the number of White males over 50 moved by raking to 159, only 0.3% of the population.

Table 4B: Estimates for White Males over 50 in Columbia

Estimate	Owner	Renter
Initial Estimate BIAS	46432 1697	5073 448
DSE before Raking (se) Undercount Rate	48317 (875) 3.90%	5670 (395) 10.53%
Rake by Tenure (se) Undercount Rate MSE / BIAS Difference from DSE	49097 (888) 5.43% 1118 / 678 +781	4889 (189) -3.76% 704 / 678 -781
Rake by Tenure, Minority Status, & +/- 50 (se) Undercount Rate MSE / BIAS Difference from DSE	48476 (878) 4.22% 868 / n/a +159	5511 (295) 7.95% 266 / n/a -159
MSE = (Est-DSE) ² - Var(Est-DSE) + Var(Est) Bias ² = (Est-DSE) ² - Var(Est-DSE)		

Table 5 summarizes the results by averaging over the 84 poststrata. In Sacramento, raking reduces the variances and the mean square errors substantially with or without the interactions, with no measurable bias in either case. In Columbia, raking reduces the estimated variances but the added bias increases the mean square error 18% overall and 30% for persons over 50. Including the interactions adds back about half of the variance reduction, but the reduced bias decreases the mean square error to a 13% increase overall and an 8% decrease for persons over 50. Including the interactions also decreases the number of persons shifted by raking, especially for those over 50. In Sacramento raking changes the 56 collapsed coverage factors by an average of 2.3%. Including the interactions decreases this change to 2.0%. In Columbia, the corresponding decrease for the 28 collapsed poststrata is from a 3.7% change to a 2.7% change; for the 8 collapsed poststrata for persons age 50 and over the decrease is from a 4.5% change to a 1.6% change. As expected, including the interactions is preserving the difference in coverage by tenure for those over or under age 50 which is lost if only tenure is included in the second dimension of the raking matrix.

Table 5: Average Statistics for Raking for two Options over 84 Poststrata (24 for over age 50)

	Sacramento	Columbia	
		Total	Over50
No Raking: \overline{VAR}	44929	356755	168830
Rake by tenure			
\overline{VAR}	30078	259479	156408
<i>moved by raking</i>	39	112	165
\overline{MSE}	16640	421955	218656
$\overline{BIAS^2}$	-13437	162477	62248
Rake by tenure, minority status, and over/under 50			
\overline{VAR}	32821	307986	161927
<i>moved by raking</i>	30	83	57
\overline{MSE}	20474	402549	155709
$\overline{BIAS^2}$	-12348	94563	-6218

The results comparing the raked estimates with or without the interactions to the unraked DSEs for the individual poststrata are displayed in the graphs at the end of the paper. Including the interactions had little effect in Sacramento where the corresponding squares and diamonds are fairly close to one another. However, in Columbia, including the interactions produces 10% differences for the White renters over age 50.

B. Additional Variables

Additional variables have been proposed for poststratification in the second dimension of the raking matrix for Census 2000. These include:

- Geographic variables. These could be for major areas such as Census regions, Census Divisions, or states, or for subareas such as urban versus nonurban areas or mailout/mailback versus update/leave areas. These variables were not applicable in Sacramento, but could have been important in the Columbia site. Investigation of these variables at the national level with the 1990 PES data can be found in Farooque (1999).
- Neighborhood characteristics such as mail return rate, percentage minority, or poverty rate. These can be used separately or combined into a short form or long form neighborhood "hard-to-count" score. Variables based on data from the Census 2000 "long form" would have to use 1990 data, while a "short form" score could use Census 2000 data. Farooque (1999) is finding that mail return and minority rates are statistically significant.
- Household composition variables which attempt to identify households and residents which are more likely to have good coverage. A simple variable with some effectiveness is whether the first two persons in a household are married. A more complex variable which is very significant in the logistic regression work in Farooque (1999) places single persons over 50 and married couples over 30 and their minor children (and then only if the only other persons in the household are older children and at most one elderly parent) in Class 1 and everyone else in Class 2. This particular variable is about as important as race/Hispanic origin or age/sex as an indicator of coverage. Unfortunately, these variables are influenced by coverage which tends to be better in the P-Sample. This results in more people being in Class 1 in the E-Sample and in Class 2 in the P-Sample. This imbalance leads to biases which cannot be eliminated. Unless a workable definition of this significant variable can be found, it probably should not be used.

IV. Conclusions

- Raking, or iterative proportional fitting, can be a valuable tool in reducing both the variance and the total error in the dual system estimation for Census 2000.
- Inclusion of interactions for race and age can control

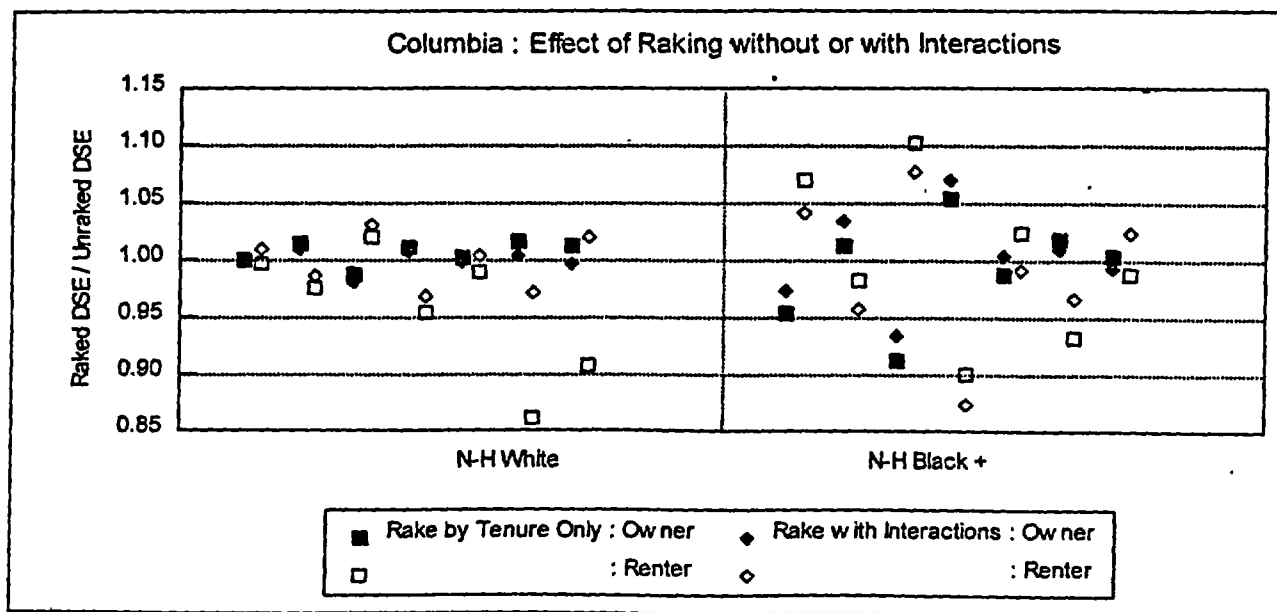
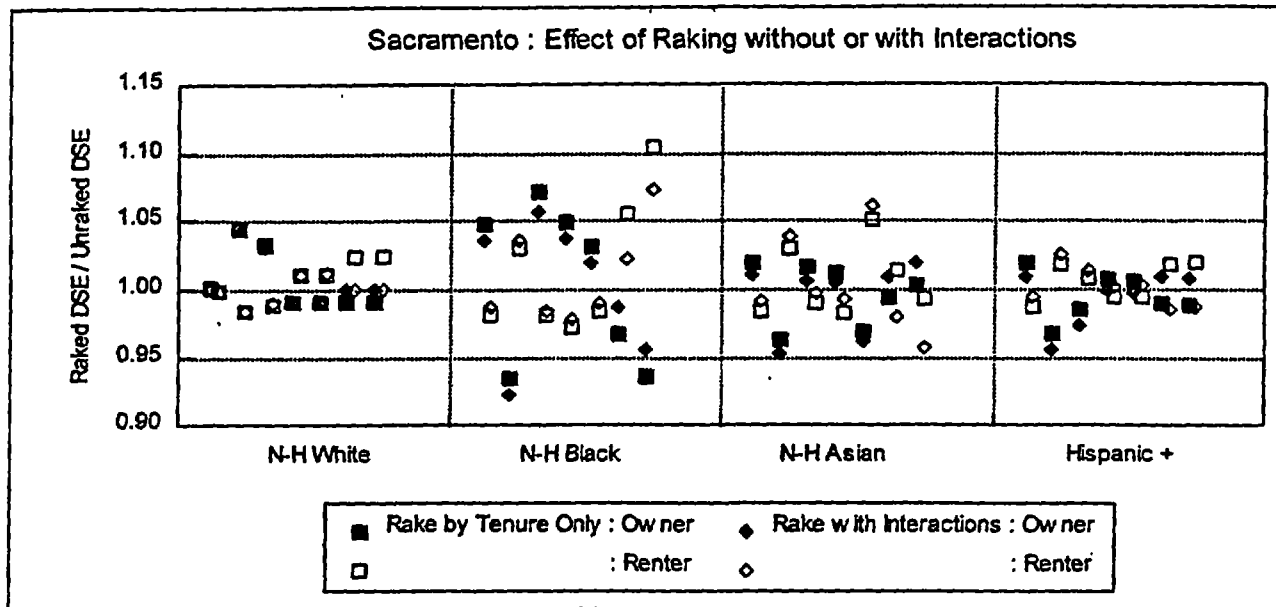
the bias due to raking when the simpler raking model is inappropriate for a particular subpopulation as occurred in Columbia.

V. References

Farooque, Golam, (1999) " Variables Selection and Determination of Raking, Dimensions," in

Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria VA, to appear.

Schindler, Eric and Griffin, Richard (1997), "2000 Census ICM: Stratification and Poststratification," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria VA, pp689-694.



VARIANCE ESTIMATION FOR THE MULTIPLICITY ESTIMATOR IN THE SERVICE BASED ENUMERATION PROGRAM

Roger Shores, Patrick J. Cantwell, and Felipe Kohn, U.S. Bureau of the Census
Roger Shores, U.S. Bureau of the Census, Washington, D.C. 20233

Key words: shelter population, soup kitchens, Bernoulli model.

discussion.

One day is selected, and everyone is counted on that day. A multiplicity estimator for the shelter component of the estimator is given by

1 Introduction

Service Based Enumeration (SBE) is the statistical program that the Census Bureau uses to estimate the population of persons without usual residence who use services. The methodology selected to measure this population is a multiplicity estimate of the number of times they use service facilities. This paper first presents the justification of the estimator and a derivation of its variance. The estimator of this variance then follows in a straightforward fashion. We examine the behavior of the multiplicity estimator and its variance. An important specific case is the one in which usage is assumed to follow a Bernoulli distribution. Results are presented that show what happens to the variance when the probability parameter for the Bernoulli distribution is varied.

2 SBE Methodology

2.1 The SBE Estimator

Multiplicity estimation is the methodology selected for use in the SBE program. Part of the population is enumerated on a specified day and asked about their use of services during a recent reference period. This information allows us to estimate the size of the total population using services. There are many multiplicity estimators, each relying on a different multiplicity rule. The SBE estimator relies on two usage questions to obtain the data. One question asks about shelter usage, while the other, directed at people who do not use shelters, asks about usage in soup kitchens and mobile food vans.

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of current research and to encourage

where n represents the number of persons enumerated at a shelter on the selected day, and A_k represents the number of days person k used a shelter during the shelter reference week, that is, the current day and the six prior days.

Enumeration at soup kitchens and mobile food vans took place the day after enumeration at shelters. The soup kitchen and mobile food van component is given by

where m represents the number of persons enumerated the next day at a soup kitchen or a mobile food van during the soup kitchen reference week, which consisted of the next day and the six prior days. B_h represents the number of days person h received a meal from a soup kitchen or mobile food van. This summation does not include people who used a shelter during the shelter reference week. The combined estimator is then

(3) For more information on this estimator see Kohn and Griffin (1999).

2.2 Justifying the Estimator Statistically

In order to justify our use of the estimator in (3), we describe circumstances under which it is appropriate. In this section and the next two, we examine the properties of the shelter-only estimator, as given in (1). The results will then be extended to the combined estimator for the shelter and soup kitchen populations in section 2.5. To

start we make the following assumptions:

- (a) The entire population of shelter users can be divided into eight mutually exclusive groups G_0, G_1, \dots, G_7 , where G_i includes all those who used shelters i times in the reference week.
- (b) Each person in G_i used the shelter on i days randomly selected during the reference week.

- (c) Users in the population visit shelters independently of each other.
- (d) There is no response error. That is, the number of days given as the frequency of shelter use during the reference week is the true number.

Obviously, these assumptions do not hold in reality. For example, consider (b). It is likely that shelters experience heavier usage certain days of the week or different times of the month. Indeed, weather may be an important factor. Assumption (c) ignores the clustering effect of companions and of mothers with children. The most questionable assumption is (d). It is likely that many users will simply not recall how many times they have visited shelters over a week's time. However, there are few inferences we can make without these or other such assumptions.

It is worth mentioning what these assumptions do *not* imply. (1) They do *not* assume that each person in the population behaves the same way with respect to the use of shelters. That is, the probability that a person falls in G_i can vary from person to person. (2) For any individual, the mechanism for determining whether a visit is made to a shelter need not be independent over the days of the week.

Note that this population does *not* include people who *never* use shelters. The goal of SBE is not to estimate all homeless people, but simply those who use shelters (and, in section 2.5, soup kitchens). The set of people in G_0 are those who sometimes use shelters but did not use a shelter during the reference week. Among shelter users, only those in G_0 can be included in the soup kitchen and mobile food van estimate. That estimate therefore covers the people in G_0 , and those who never use shelters but sometimes use soup kitchens and mobile food vans.

Let N_i be the number of people in group G_i , and let N be the size of the shelter population, that is,

From equation (1) one sees that the shelter-only multiplicity estimator is the sum of 7 over A_k , where A_k is the number of shelter visits made during the reference week for the k th person, over the population enumerated at shelters. By grouping together the A_k 's corresponding to people with the same number of visits (that is, into their groups G_i), we can rewrite the estimator as where, for $i = 1, 2, \dots, 7$, n_i is the number of people counted in the SBE operation (out of N_i people in G_i) who visited shelters i times in the past week.

It is easy to determine the conditional distribution of the n_i . Under the assumptions above, given N_i , n_i follows a binomial distribution with N_i trials and probability of success equal to $i/7$. (One observes that n_7 is equal to N_7 with certainty).

Since n_i is an obvious estimator for N_i , the shelter-only multiplicity estimator, \hat{N} , can be defined as

. We see an immediate problem: with the shelter-only estimator, \hat{N} , we have not included a component to estimate N_0 , the number of people in G_0 . We leave the derivation of more complex methods to estimate N_0 to another paper, but present a partial remedy in section 2.5, where we investigate the combined estimator for shelters and soup kitchens.

2.3 Conditional Mean and Variance of

Because each n_i follows a binomial distribution with parameters N_i and $i/7$, the derivation of the expected value and variance of the multiplicity conditional estimator is straightforward:

$$E(n_i | \{N_0, N_1, \dots, N_7\}) = N_i (i/7), \text{ and}$$

$$E(\hat{N} | \{N_0, N_1, \dots, N_7\}) = N - N_0.$$

Clearly, \hat{N} is biased downwards by the amount N_0 .

Under assumptions (a) and (c), and conditional on the set $\{N_0, N_1, \dots, N_7\}$, the random variables n_1, n_2, \dots, n_7 are stochastically independent. It then follows that

$$\begin{aligned} \text{Var}(\hat{N} | \{N_0, N_1, \dots, N_7\}) &= \text{Var}(n_i | \{N_0, N_1, \dots, N_7\}) (7/i)^2 \\ &= N_i (i/7) (1 - i/7) (7/i)^2 \\ &= N_i (7-i) / i \end{aligned} \tag{4}$$

We obtain a straightforward estimator for this variance by estimating the population components N_i :

$$\begin{aligned} \text{Var}(\hat{N} | \{N_0, N_1, \dots, N_7\}) &= n_i (7/i) (7-i) / i \\ &= n_i 7 (7-i) / i^2. \end{aligned} \tag{5}$$

Finally, we can obtain the conditional variance of this variance estimator:

$$\begin{aligned} \text{Var} ((1 \{N_0, N_1, \dots, N_7\} | \{N_0, N_1, \dots, N_7\})) \\ = N_i (i/7) (1 - i/7) [7 (7-i) / i^2]^2 \\ = N_i (7-i)^3 / i^3. \end{aligned} \quad (6)$$

As one would expect, for a fixed total number of people $N - N_0$, the true variance increases as the people make fewer visits to shelters during the reference week. In that case, fewer shelter people tend to be enumerated, and the weights applied to their records, $7/i$, tend to be larger.

A simple estimator for the variance of $(1 \{N_0, N_1, \dots, N_7\})$ in (6) can be given by inserting estimates, $n_i (7/i)$, for each N_i .

2.4 Unconditional mean and variance

The results in the previous section demonstrate what happens conditional on the population sizes N_1, N_2, \dots, N_7 . But the behavior of the estimator also depends on the stochastic mechanism that produces the values of the N_i 's. After all, someone who visits a shelter four times one week may well visit twice in another week. How does this variability affect the distribution of?

The unconditional mean of is as follows:

$$E() = E(E(n_i (7/i) | \{N_0, N_1, \dots, N_7\})),$$

where the outside expectation is taken over the distribution of possible values of the vector $\{N_0, N_1, \dots, N_7\}$. As the terms $n_i (7/i)$ are conditionally unbiased for the N_i , we can write

$$\begin{aligned} E() &= E[N_i] = E[N - N_0] \\ &= N - E[N_0]. \end{aligned}$$

Thus the mean of has a downward bias equal to the expected value of N_0 , the expected size of the shelter population who do not make a visit in the reference week (G_0). The unconditional variance can be derived similarly:

$$\begin{aligned} \text{Var}() &= \text{Var}[E(n_i (7/i) | \{N_0, N_1, \dots, N_7\})] \\ &\quad + E[\text{Var}(n_i (7/i) | \{N_0, N_1, \dots, N_7\})] \\ &= \text{Var}[N_i] + E[N_i (7-i) / i] \end{aligned}$$

$$= \text{Var}[N - N_0] + E(N_i) (7-i) / i$$

$$= \text{Var}[N_0] + E(N_i) (7-i) / i. \quad (7)$$

The leading component of the variance in (7), $\text{Var}[N_0]$, may be small if the term N_0 tends to be small. Note that the result in (7) requires only the assumptions made in section 2.2. It is not necessary that different people in the population visit shelters in the same way or with the same probabilities.

2.5 Extending to the combined shelter and soup kitchen estimator

To this point we have investigated the distribution of the shelter-only estimator. The extension to the combined shelter and soup kitchen multiplicity estimator is important because we can account for (i) people who visit shelters at times, but did not during the reference week, and (ii) people who never frequent shelters, but who sometimes visit soup kitchens.

Let us recall the procedure applied at *soup kitchens*: all people enumerated there are asked two questions: (1) how many times in the reference week they visited a soup kitchen, and (2) whether they visited a shelter at any time during the reference week. If the answer to (2) is "yes," their response does not contribute to the soup-kitchen component of the combined estimator, as they were *represented in the shelter population* already--whether or not they were enumerated at a shelter. The idea is to represent each person, that is, give him or her a chance to be enumerated, *but only once*. If the answer to (2) is "no," they are assigned a weight equal to seven times the reciprocal of their response to (1), similar to the weighting used in the shelter-only estimator. The result is the combined estimator in equation (3).

To justify the combined estimator, we first define the population more generally than before. We include all people who sometimes use a shelter *or* a soup kitchen *or both*, although they may have used neither in the reference week leading to enumeration day. This population can be divided into 64 groups G_{ij} of size N_{ij} , where G_{ij} includes all those who visited a shelter i times during the reference week, and also visited a soup kitchen j times during its week. In the groups G_{0j} , $j = 0, 1, \dots, 7$, are those people who did not use a shelter during the reference week, including those who sometimes use a shelter and those who never do. According to our definition of the population, those in G_{00} sometimes use a shelter or a soup kitchen.

Because we exclude from the soup-kitchen component of the estimator any census respondents who

used a shelter in the reference week, we can limit ourselves to estimating the following components of N : $N_1, N_2, \dots, N_7, N_{01}, N_{02}, \dots, \text{and } N_{07}$. All persons in G_{ij} , where $i > 0$, *do not* contribute toward the second summation in (3), the soup-kitchen component, *but are represented* in the estimation of N_i through the first component--whether or not they were enumerated at a shelter.

We extend the assumptions given above in section 2.2 analogously to the use of soup kitchens. Then the combined estimator can be written as

$$= n_i (7/i) + n_{0j} (7/j),$$

where the n_{0j} are the number of people in G_{0j} enumerated in the SBE operation at soup kitchens. The results derived in the previous sections carry forward analogously. Conditional on the set of population sizes $\{N_1, N_2, \dots, N_7; N_{01}, N_{02}, \dots, N_{07}\}$,

$$E(\mid \{N_1, N_2, \dots, N_7; N_{01}, N_{02}, \dots, N_{07}\})$$

$$= N_i + N_{0j} = N - N_{00}, \text{ and}$$

$$\text{Var} (\mid \{N_1, N_2, \dots, N_7; N_{01}, N_{02}, \dots, N_{07}\})$$

$$= N_i (7-i) / i + N_{0j} (7-j) / j.$$

This suggests the following conditional estimator of the variance of N hat:

$$(\mid \{N_1, N_2, \dots, N_7; N_{01}, N_{02}, \dots, N_{07}\})$$

$$= n_i 7 (7-i) / i^2 + n_{0j} 7 (7-j) / j^2. \quad (8)$$

As with the shelter-only estimator, the conditional variance of the variance estimator is easily obtained:

$$\text{Var} (() \mid \{N_1, N_2, \dots, N_7; N_{01}, N_{02}, \dots, N_{07}\})$$

$$= N_i (7-i)^3 / i^3 + N_{0j} (7-j)^3 / j^3 \quad (9)$$

Unconditionally, the mean and variance of are

$$E () = E [N_i + N_{0j}] = N - E [N_{00}],$$

and

$$\begin{aligned} \text{Var} () &= \text{Var} [N_{00}] + E(N_i) (7-i) / i \\ &\quad + E(N_{0j}) (7-j) / j. \end{aligned} \quad (10)$$

Again we see that the estimator is biased downward, now by $E [N_{00}]$, the mean number of people who

frequent shelters or soup kitchens, but who visit *neither* during the appropriate reference week. This number is generally smaller than $E [N_0]$, the bias in the shelter-only estimator. If we are willing to apply information about the behavior of people in the target population, we can model the distribution of N_{00} and the set $\{N_1, N_2, \dots, N_7; N_{01}, N_{02}, \dots, N_{07}\}$, and thereby predict the unconditional performance of the combined estimator. In fact, under reasonable assumptions the unconditional mean and variance of N_{00} may contribute only a very small part of the total mean and variance.

3 Constant Visit Probabilities Over the Population

To investigate the behavior of the multiplicity estimator under actual conditions, we add one assumption to those made in section 2.2:

- (e) The probability, f_{ij} , $i = 0, 1, \dots, 7$, that a person makes i visits to a shelter and j visits to a soup kitchen during the respective reference weeks, is the same for each person.

Note that we are not yet making any further claims about the day-to-day behavior of the individuals beyond what has already been assumed in section 2.2.

Consider first the shelter-only estimator. Analogous to previous notation, for $i = 0, 1, \dots, 7$, let f_i be the sum of the f_{ij} 's as j runs from 0 to 7. Under (e) and the prior assumption of independent shelter use ((c) in 2.2), for a population of N people who sometimes use shelters, the set $\{N_0, N_1, \dots, N_7\}$ follows a multinomial distribution with N trials and probability parameters f_0, f_1, \dots, f_7 . Thus each N_i has a binomial distribution with parameters N and f_i . The conditional results of section 2.3 pertaining to the multiplicity estimator continue to hold. But now we can derive its unconditional mean and variance as well. Substituting into the equations in section 2.4, we conclude that

$$E () = (N - N \cdot f_0) = N (1 - f_0), \text{ and}$$

$$\text{Var} () = N f_0 (1 - f_0) + N f_i (7-i) / i.$$

If f_0 is small, the relative bias is small, and the first component of the unconditional variance is small relative to the other components.

Extending these results to soup-kitchen enumeration is straightforward. The number in the population who make no visits to shelters or soup kitchens during the reference weeks, N_{00} , follows the binomial law with parameters N and f_{00} . The bias in the combined estimator is $E [N_{00}] = N \cdot f_{00}$, and the variance is

$$\text{Var}() = N f_{00}(1 - f_{00}) + N f_i(7-i)/i + N f_0(7-j)/j. \quad (11)$$

4 The Bernoulli Model for Visit Behavior

In this section we discuss the case where individual service usage follows a Bernoulli distribution. We make one last assumption:

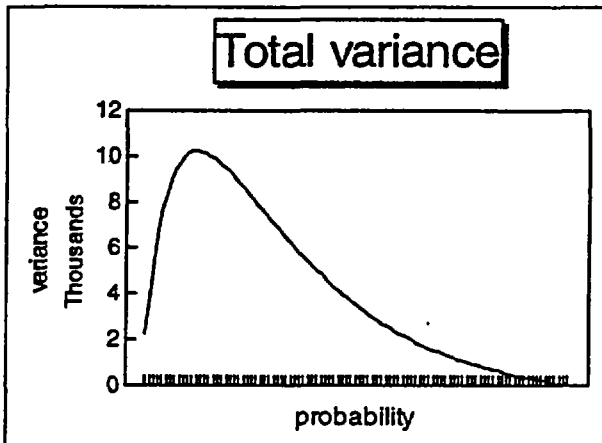
- (f) On any day in the reference week, each person is assumed to use a shelter (soup kitchen) with probability p_1 (p_2), with behavior independent from day to day and over facilities.

Thus for any person the number of visits to a shelter (soup kitchen) over the seven days is binomially distributed with parameters 7 and p_1 (p_2); the f_i 's and f_0 's above are replaced by values of the binomial probability function. It follows easily that

$$E() = N(1 - f_{00}) = N[1 - (1-p_1)^7(1-p_2)^7],$$

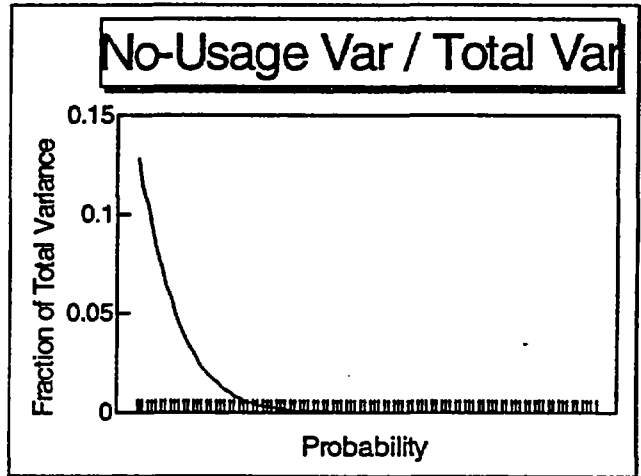
and from equation (11),

$$\text{Var}() =$$



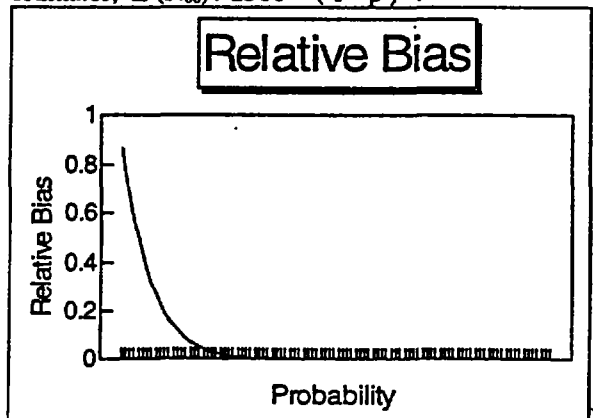
We will make the simplifying assumption that usage at shelters and soup kitchens occurs with the same probability; that is, $p_1 = p_2 = p$. Different probabilities of service usage can produce very different estimates of the unconditional variance and of related statistics. These statistics are proportional to N , but it is illustrative to hold N constant while varying p . The following graphs give the variance and other statistics as a function of p , while holding N constant at 2,500.

The first graph gives the total unconditional variance; it reaches a maximum of about 10,200 at $p = 0.13$. As p is reduced from 1, the variance at first increases because the counts in the lower usage categories for the shelter and soup kitchen variances will increase, and, as we have seen, the lower usage categories have more of an effect on the variance than do the higher usage categories. Eventually, however, as p is reduced further, so many people fall into the no-usage category that the variance of the other two components begins to decrease, so that the total variance will decrease.

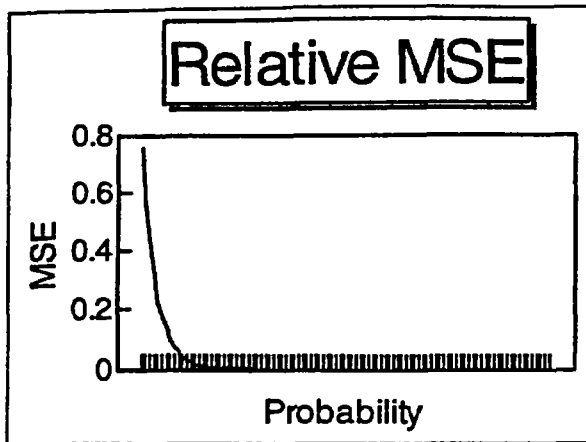


The no-usage variance also decreases once p becomes small enough. It does, however, assume a larger fraction of the total as p becomes smaller. The next graph presents that variance as a fraction of the total variance. It shows that the no-usage component of the variance is only significant when p is small. At $p = 0.13$, it is only about 3% of the total, and at $p = 0.20$, it has already fallen to approximately 1%.

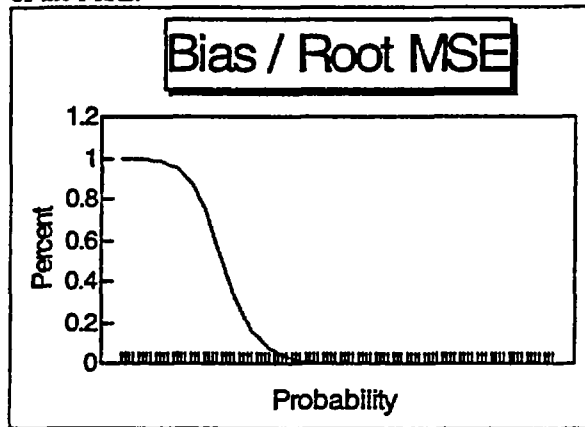
The next graph shows the relative bias of the estimator, $E(N_{00}) / 2500 = (1-p)^{14}$.



It too quickly falls to insignificance as p increases, equaling, for example, approximately 0.02 at $p = 0.25$.



Next, we see a graph of the relative mean square error. Finally, we see the bias as a fraction of the square root of the MSE.



The relative bias is initially substantial. At $p = 0.01$, for example, it equals 87%. As one can see from the prior graph, it falls quickly; as long as p is equal to at least 0.3, it will be under 1%. The relative mean square error declines in a similar manner. Thus the bias is, at the lowest usage probabilities, a substantial fraction of the root MSE, making up essentially all of it for probabilities of about 0.1 or less. Once p has risen to 0.3, however, this fraction has fallen to approximately 23% and is declining rapidly as a proportion of the

total. By $p = 0.4$ it contributes only about 3%, and at $p = 0.5$ it has become negligible as a fraction of the root MSE.

To make this analysis more realistic, we can let p_1 differ from p_2 . The general results will be similar, but we can examine the effect of differences in usage at shelters and soup kitchens. Such results are not included here because of limited space.

5 Continuing Research

Currently we are pursuing several areas of research:

- Estimating the "no-usage" component of N , that is, N_0 or N_{00} , from the enumeration data.
- Modeling the behavior of the individuals in the population allowing for different probabilities of service usage from person to person.
- Determining the distribution of $\{N_0, N_1, \dots, N_7\}$ under various behavior models.
- Approximating the actual distribution of $\{N_0, N_1, \dots, N_7\}$ by similar distributions whose parameters can be reasonably estimated.
- Evaluating the statistical properties of the derived estimators, that is, \hat{N} and \hat{N}_0 , by using simulations.

References

Kohn, F. and Griffin, R., (1999). "Multiplicity Estimators Applied in the Service Based Enumeration Program," *Proceedings of the Section on Survey Research Methods, American Statistical Association (to appear)*.

Integrated Coverage Measurement Persons Not Matched in the Census 2000 Dress Rehearsal

Glenn Wolfgang, Danny Childers, Bureau of the Census
Glenn Wolfgang, Bureau of the Census, Washington, DC 20233¹

Key Words: Census Coverage; Dual System Estimation

I. Introduction

Census 2000 procedures were rehearsed in three sites during 1998: Sacramento, California; Menominee county, Wisconsin, including the Menominee Indian Reservation; and Columbia, South Carolina with eleven surrounding counties. These sites provided responses to the initial census as well as to an independent second enumeration, the Integrated Coverage Measurement (ICM) survey, of sampled areas in these sites. (In South Carolina, the coverage measurement survey was designed as an evaluation not strictly integrated with the census. Yet, to simplify descriptions, this paper refers to all 1998 coverage measurement as ICM.)

The ICM involves two samples. Both are based on housing units found in blocks (or subsampled sections of large blocks) selected from census sites. The E-sample is comprised of persons reported in the census as Census Day residents in those housing units; it is used to count errors among those the census did count. The P-sample includes persons the survey found residing in housing units in the same areas on Census Day; it is used to determine who was missed in the census. Names of persons in the P-sample of the ICM are sought among census names. Persons found in both are matches.

The focus of this paper is on P-sample nonmatches, persons who were not found to be enumerated in the Census 2000 Dress Rehearsal. The aim is to identify characteristics that may be related to their being missed in census enumeration. The statistic used in this study is the nonmatch rate, the proportion of nonmatches among P-sample persons, computed within age, race, and other descriptive categories. The nonmatch rate is well related to (but less refined and more inflated than) the dual system adjustment factor used in census coverage evaluation. Errors and incomplete data estimated from the E-sample, as well as matches that may exist among census enumerations beyond areas searched, are refinements taken into account by dual system estimates but not nonmatch rates. Nonmatch rates are worthy of study independent of the effects of false or ambiguous

enumerations, which are investigated by Feldpausch and Childers (1999) and by Jones and Childers (1999). Beaghen (1999) modeled both E-sample and P-sample data to gain insight into misses.

Several prior publications provide more background for this research. The procedures for both census and survey data collections in the Census 2000 Dress Rehearsal are described in Waite and Hogan (1998). Childers (1999) describes the ICM more completely. P-sample nonmatches were analyzed in the context of the 1990 census results by Moriarity and Childers (1993) using some of the same variables investigated in this research.

II. Limitations

This paper has a specific focus on P-sample nonmatches and the missed enumerations they represent. That excludes some other operational and theoretical concerns worthy of study. Other issues or errors beyond the scope of this paper include:

- Error in enumerations made by the census, measured with the E-sample,
- Imputation error in correcting ambiguity or inconsistency in census or survey data,
- Error in coding residence status or match status for the P-sample,
- Error due to whole-household nonresponse and the non-interview adjustment,
- Response error in reported characteristics,
- Correlation bias or lack of independence between census and survey enumerations, resulting in understated or overstated nonmatch rates,
- Lack of independence among groups compared or correlation due to the design,
- Incomplete representation of all areas of the United States during a decennial census. Three or so sites do not represent the variety in the nation. Dress rehearsal results also did not benefit from a full census publicity campaign.
- Interactions among variables within site.

¹This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

III. Methods

This study used final person-level data from the 1998 Dress Rehearsal for Census 2000. More detail on the processing of the data may be found in Waite and Hogan (1998). Here is a brief overview of the matching work. The P-sample people and the census people from the Census Unedited File (CUF) were computer matched within cluster. The computer matching involved first standardizing the name formats. Names and person characteristics of the P-sample people were compared to those of census people. A ranking score was assigned to each pair of person records and the optimal pairings were identified. Those pairs were reviewed to determine cutoffs in the scores taken to separate matches, possible matches, and nonmatches. Match cutoffs are assigned conservatively so there are virtually no false matches.

The possible matches and P-sample nonmatches were clerically reviewed using an automated match and review system. The names, age, race, Hispanic origin, sex, relationship, family composition, and address are displayed for review by the matching clerks, who matched some people the computer could not. After the matching, field follow-up was conducted to resolve or confirm coding of selected cases.

If match status remained unresolved, match probability was imputed. When variables used for poststratification (age, sex, race, ethnicity, and tenure) were missing, values were imputed. Each record was given a sampling weight derived from the probability of selection of the block cluster and of the block segment within the block if it was very large. For households not successfully interviewed, a non-interview adjustment was applied to similar other households. Sampling weights and non-interview adjustments were applied in all analyses.

A nonmatch rate, the weighted number of nonmatches divided by the weighted number of P-sample persons, was computed for various groups within the P-sample. Identifying groups with unusually high nonmatch rates provides insights on conditions associated with missed census enumerations. For this purpose, P-sample persons are grouped using variables:

- Site – The Dress Rehearsal for Census 2000 was conducted within three states. South Carolina was so diverse that it was useful to divide it into three parts and analyze each as if it were a separate site. The resulting five sites vary on urban-rural character and type of enumeration area (TEA). TEA= 1, mail-out/mail-back, is the method of data collection used when the census address is specific enough to ensure the form can be delivered to only one location. TEA= 2, update/leave, means each address in sample will be updated in the field and a census form left for mail return:
 - S, CA = Sacramento, CA – Urban; TEA = 1,
 - C, SC = Columbia, SC – Urban; TEA = 1,
 - O, SC = Other SC – Mixed; TEA = 1,
 - R, SC = Rural SC – Rural; TEA = 2,
 - M, WI = Menominee, WI – Rural; TEA = 2.
- Race and ethnicity – Respondents were asked to identify ethnicity (Hispanic or non-Hispanic) as well as any race that applied (Black, American Indian, Asian, Pacific Islander, White, or Other). A person with multiple race responses was assigned to the largest group, other than white, based on the site's 1990 census data. Groups were collapsed, as needed in a site, to form the poststrata used in final dual system estimates. Schindler (1999) describes race categories more fully. The Sacramento site had four:
 - Hispanic, American Indian, and Pacific Islander
 - Black (non-Hispanic)
 - Asian (non-Hispanic)
 - White (including "other"; non-Hispanic)In South Carolina and Menominee, Black, Hispanic, American Indian, Asian, and Pacific Islander groups were collapsed into one group, Blacks predominate in South Carolina; American Indians in Menominee.
- Sex (Male or Female)
- Age – Categories of age over age 17 are usually split by gender forming seven age/sex poststrata. In this paper, only basic age groups are analyzed:
 - Under age 18
 - Age 18 to 29
 - Age 30 to 49
 - Age 50 or more
- Tenure (Home owner or Renter)
- Impute – Was age, sex, race, ethnicity, or tenure imputed for the ICM? (Imputed or not)
- Proxy – (Proxy respondent or household member)
- Mover – Did the census day resident move from the sample address before the survey interview? (Mover or Nonmover)
- Household Size – The number of non-imputed P-sample persons enumerated at the address:
 - One (enumerated on one census form)
 - 2-5 (all could be enumerated on one census form)
 - 6 or more (supplemental forms required)
- Household Structure – Type of structure at the address:
 - Single dwelling
 - Multi-unit
 - Mobile home
- Address Style – Addresses may be written in various styles. Some are more useful or, at any rate, easier than others in distinguishing one housing unit from another or pinpointing its location on a map.
 - Street -- House number and street name,
 - Rural route -- Rural route number or street name,

- PO Box – Post office box number
- Subsampling – Was there any subsampling within the block cluster (Subsampled or not)
- Growth – Amount of growth in the number of addresses for the area (tract) since 1990
 - Less than 20%
 - 20% to 40%
 - 40% or more
- Relisting – To insure the quality of address lists in a sample block cluster, field staff sometimes revisited the area and recreated the listing. The conditions leading to relisting might be related to those causing census misses (Relisted or not)

The group nonmatch rates were compared using VPLX, software designed to estimate variances in complex sample surveys using replication methods. VPLX was developed by Bob Fay, Senior Mathematical Statistician at the U.S. Bureau of the Census, and described in Fay (1990). VPLX documentation and software are available at www.census.gov/sdms/www/vwelcome.html. Stratified Jackknife methods were used to compute variance estimates for the nonmatch rates. VPLX also generated t-tests among these rates.

The tests were conducted at a 90% level of confidence, using a multiple comparison of means technique with a Bonferroni criterion, as described by Games (1971). It controls the probability of Type I error for a family of tests. In the context of this paper, a family of tests is defined as all tests conducted between groups of cases that together comprise the whole site or, when comparing sites, the whole sample. For example, when comparing the four race groups within the Sacramento site, six pairs of nonmatch rates were tested. To control the chance of Type I error at $\alpha = 0.10$ for all six tests combined, we used an adjusted criterion t-value associated with the probability of one of six two-tailed tests that have a joint error probability equal to 0.10. In addition, tests with groups based on less than 100 person records were avoided, either through collapsing with other groups or simply by dropping the group from that family of tests.

IV. Results

In general, results are presented in tables displaying group names (and group numbers among groups being compared, usually within site) nonmatch rates (rate), the stratified jackknife standard error (s. e.), the number of persons contributing data to the analysis (n), and a list of the numbers of groups with which a significant difference was found (*). Criterion t-values (e.g., $|t| > 1.65$) vary, as described above, with the number of comparisons being made in the family of tests. They are displayed below each table. The groups are arranged by nonmatch rate from lowest to highest to help display data patterns.

Overall, the nonmatch rates in these dress rehearsal sites are higher than a corresponding rate based on dual system estimates qualified by erroneous enumerations and insufficient information. They are also higher than 1990 rates. That is partly due to differences in how often and how far into blocks surrounding the sample block the search for census matches and counterbalancing erroneous enumerations was extended. In 1990, the search area was one ring of surrounding blocks in urban areas and two rings of blocks in rural areas. In dress rehearsal, the search area was the sample block cluster, except in a small number of clusters. The dress rehearsal did not have all the coverage improvement activities of a decennial census, such as coverage edit follow-up and block canvassing. Such efforts increase coverage.

The first table shows that the sites, picked to test a variety of census collection conditions in the U.S., did indeed differ from each other, but not when grouped by TEA, which usually yields differing nonmatch rates (Moriarty and Childers, 1993).

Table 1: Nonmatch Rates by Census 2000 Dress Rehearsal Site

SITE	rate	*	s. e.	n
1. C, SC	0.154	3-5	0.007	17810
2. M, WI	0.171	3-5	0.015	1271
3. S, CA	0.218	all	0.004	36336
4. R, SC	0.255	1-3	0.012	5359
5. O, SC	0.280	1-3	0.008	12751

* groups differing at $|t| > 2.57$

The results of other analyses are presented separately within sites. Tables 2, 3, 4, and 5 show comparisons of groups defined by major poststratification variables, which historically yield differences in coverage rates, thus making the poststratification useful in improving the dual system estimates. Poststratum groups compared below differ in a few instances from those used in official dual system estimates, mainly in order to arrange comparison groups with sample sizes greater than 100.

Table 2: Nonmatch Rates by Site and Tenure

SITE	TENURE	rate	*	s. e.	n
S, CA	1. Owner	0.161	2	0.004	19613
	2. Renter	0.284	1	0.006	16723
C, SC	1. Owner	0.121	2	0.005	8459
	2. Renter	0.174	1	0.011	9351
O, SC	1. Owner	0.261	2	0.009	9436
	2. Renter	0.333	1	0.018	3315
R, SC	1. Owner	0.243	2	0.013	4549
	2. Renter	0.316	1	0.032	810
M, WI	1. Owner	0.162	.	0.019	903
	2. Renter	0.187	.	0.023	368

* groups differing at $|t| > 1.65$

These poststrata results were consistent with 1990. Nonmatch rates for renters were higher than for owners, except in Menominee. Whites generally had the lowest nonmatch rates. Asian and Hispanic race/ethnicity groups shared nearly equal, moderate rates in Sacramento. Those age 50 or over, and next those aged 30-49, had the lowest rates and young adults (aged 18-29) the highest, except in Menominee. Females often had lower nonmatch rates than males. Although the trends were generally consistent from site to site, significant differences tended to appear more in sites with larger sample sizes.

Table 3: Nonmatch Rates by Site and Race/Ethnicity

SITE	RACE	rate	*	s. e.	n
S, CA	1. White	0.185	all	0.005	17038
	2. Asian	0.229	1,4	0.012	5668
	3. Hisp.	0.233	1,4	0.007	8105
	4. Black	0.286	all	0.010	5525
C, SC	1. White	0.116	2	0.006	9563
	2. Black	0.192	1	0.010	8247
O, SC	1. White	0.246	2	0.008	7329
	2. Black	0.325	1	0.015	5422
R, SC	1. White	0.197	2	0.011	3330
	2. Black	0.352	1	0.024	2029
M, WI	1. White	0.105	2	0.030	269
	2.Am.Ind.	0.184	1	0.015	1002

* groups differing at $|t| > 2.39$ in S, CA
or $|t| > 1.65$ elsewhere

Table 4: Nonmatch Rates by Site and Age

SITE	AGE	rate	*	s. e.	n
S, CA	1. 50 +	0.152	all	0.007	9090
	2. 30-49	0.195	all	0.005	11156
	3. 1-17	0.265	1,2	0.009	9852
	4. 18-29	0.279	1,2	0.008	6238
C, SC	1. 50 +	0.104	all	0.007	4246
	2. 30-49	0.154	1	0.009	5721
	3. 1-17	0.171	1	0.023	4143
	4. 18-29	0.182	1	0.011	3700
O, SC	1. 50 +	0.233	3,4	0.012	3592
	2. 30-49	0.269	4	0.011	3919
	3. 1-17	0.310	1	0.023	3239
	4. 18-29	0.338	1,2	0.016	2001
R, SC	1. 50 +	0.193	3,4	0.017	1467
	2. 30-49	0.249	.	0.019	1722
	3. 1-17	0.278	1	0.029	1360
	4. 18-29	0.337	1	0.033	810
M, WI	1. 50 +	0.098	4	0.022	372
	2. 30-49	0.166	.	0.024	311
	3. 18-29	0.199	.	0.039	161
	4. 1-17	0.216	1	0.025	427

* groups differing at $|t| > 2.39$

Table 5: Nonmatch Rates by Site and Sex

SITE	SEX	rate	*	s. e.	n
S, CA	1. Female	0.210	2	0.005	18823
	2. Male	0.226	1	0.005	17513
C, SC	1. Female	0.141	2	0.008	9608
	2. Male	0.170	1	0.008	8202
O, SC	1. Female	0.279	.	0.010	6790
	2. Male	0.281	.	0.010	5961
R, SC	1. Female	0.245	.	0.015	2833
	2. Male	0.266	.	0.016	2526
M, WI	1. Female	0.164	.	0.021	654
	2. Male	0.179	.	0.020	617

* groups differing at $|t| > 1.65$

There was little imputation of poststratification variables, but when there was enough to test, imputed data had higher nonmatch rates. Households with proxy respondents yielded higher nonmatch rates than those with household member respondents (HHR), as in 1990.

Table 6: Nonmatch Rates by Site and Imputation

SITE	IMPUTED	rate	*	s. e.	n
S, CA	1. No imp.	0.217	2	0.004	35719
	2. Imputed	0.280	1	0.019	617
C, SC	1. No imp.	0.153	2	0.007	17539
	2. Imputed	0.230	1	0.034	271
O, SC	1. No imp.	0.279	2	0.008	12560
	2. Imputed	0.376	1	0.042	191

* groups differing at $|t| > 1.65$

Table 7: Nonmatch Rates by Site and Proxy Status

SITE	PROXY	rate	*	s. e.	n
S, CA	1. HHR	0.214	2	0.004	34581
	2. Proxy	0.296	1	0.013	1755
C, SC	1. HHR	0.150	2	0.007	16787
	2. Proxy	0.200	1	0.016	1023
O, SC	1. HHR	0.276	2	0.008	12146
	2. Proxy	0.377	1	0.027	605
R, SC	1. HHR	0.252	2	0.012	5196
	2. Proxy	0.364	1	0.041	163

* groups differing at $|t| > 1.65$

Movers and people with unresolved mover status had higher nonmatch rates than nonmovers. Those living in households large enough to require additional, supplemental census forms (to list the persons who did not fit on the first form) had higher nonmatch rates than those in smaller households, much as in 1990. Except in Columbia, SC, residents of single unit dwellings had lower nonmatch rates than respondents in multiple housing unit structures (buildings) or mobile homes, again, as in 1990.

Table 8: Nonmatch Rates by Site and Mover Status

SITE	MOVER	rate	*	s. e.	n
S, CA	1 Nonmover	0.209	all	0.004	34090
	2. Unresolv	0.353	1	0.022	634
	3. Mover	0.371	1	0.015	1612
C, SC	1. Nonmover	0.144	all	0.007	16478
	2. Mover	0.254	1	0.017	1064
	3. Unresolv	0.327	1	0.037	268
O, SC	1 Nonmover	0.276	all	0.008	12114
	2. Unresolv	0.364	1	0.038	209
	3. Mover	0.365	1	0.026	428
R, SC	1. Nonmover	0.251	2	0.012	5204
	2. Mover	0.431	1	0.070	139

* groups differing at $|t| > 2.13$ (or 1.65 in R,SC)

Table 9: Nonmatch Rates by Site and Household Size

SITE	SIZE	rate	*	s. e.	n
S, CA	1. 2-5	0.201	3	0.004	25072
	2. 1	0.215	3	0.008	5229
	3. 6+	0.289	1.2	0.010	6035
C, SC	1. 1	0.147	3	0.008	1046
	2. 2-5	0.148	3	0.007	13540
	3. 6+	0.257	1.2	0.025	3224
O, SC	1. 2-5	0.266	3	0.008	10230
	2. 1	0.299	3	0.016	1518
	3. 6+	0.389	1.2	0.027	1003
R, SC	1. 2-5	0.238	3	0.013	4470
	2. 1	0.284	3	0.030	467
	3. 6+	0.413	1.2	0.035	422
M, WI	1. 6+	0.151	.	0.026	343
	2. 2-5	0.179	.	0.021	844

* groups differing at $|t| > 2.13$ (or 1.65 in M, WI)

Table 10: Nonmatch Rates by Site and Type of Structure

SITE	Structure	rate	*	s. e.	n
S, CA	1. Single	0.193	all	0.004	30034
	2. Multi	0.319	all	0.011	6111
	3. Mobile	0.597	all	0.073	184
C, SC	1. Single	0.144	.	0.005	12207
	2. Multi	0.165	.	0.013	5540
O, SC	1. Single	0.253	all	0.008	9852
	2. Multi	0.326	1	0.032	1078
	3. Mobile	0.387	1	0.024	1811
R, SC	1. Single	0.215	2	0.012	3663
	2. Mobile	0.344	1	0.021	1664
M, WI	1. Single	0.152	2	0.016	1012
	2. Mobile	0.264	1	0.028	241

* groups differing at $|t| > 2.13$ in S, CA & O, SC
or $|t| > 1.65$ elsewhere

Address style groups seldom had sufficient data for a good comparison. Those in Rural S.C. with a full street address, including house number, had a lower nonmatch rate. Having a P.O. box did not differ from having no

P.O. box, whether or not other addresses were given.

Table 11: Nonmatch Rates by Site and Address Style: Full Street Address vs Part or None

SITE	Street	rate	*	s. e.	n
R, SC	1. Full	0.235	2	0.012	3499
	2. Part	0.289	1	0.024	1755
M, WI	1. Part	0.152	.	0.021	427
	2. Full	0.184	.	0.019	844

* groups differing at $|t| > 1.65$

Table 12: Nonmatch Rates by Site and Address Style: P.O. Box Address vs No P.O. Box

SITE	P.O. Box	rate	*	s. e.	n
R, SC	1. Have	0.237	.	0.033	324
	2. None	0.256	.	0.021	4930
M, WI	1. Have	0.153	.	0.021	650
	2. None	0.196	.	0.024	621

* groups differing at $|t| > 1.65$

Lower nonmatch rates were found in two sites for persons living in blocks with so many residents that the block was subsampled. The two city sites did not agree.

Table 13: Nonmatch Rates by Site and Whether Subsampled:

SITE	Subsam	rate	*	s. e.	n
S, CA	1. Yes	0.204	2	0.007	93081
	2. No	0.222	1	0.005	26678
C, SC	1. No	0.151	.	0.005	12948
	2. Yes	0.158	.	0.013	4760
O, SC	1. Yes	0.249	2	0.019	2705
	2. No	0.291	1	0.009	10016
R, SC	1. No	0.250	.	0.013	4705
	2. Yes	0.287	.	0.018	616

* groups differing at $|t| > 1.65$

Results concerning blocks that had high rates of housing construction or growth were also inconsistent. It may be that low growth areas are associated with higher nonmatch rates, but that needs further study.

Table 14: Nonmatch Rates by Site and Locality's Level of Growth:

SITE	Growth	rate	*	s. e.	n
C, SC	1. 20-40%	0.143	.	0.015	3105
	2. Low	0.158	.	0.007	14681
O, SC	1. 20-40%	0.222	all	0.024	711
	2. Low	0.281	1	0.009	10962
	3. > 40%	0.308	1	0.023	1078
R, SC	1. > 40%	0.188	3	0.029	254
	2. 20-40%	0.206	3	0.018	990
	3. Low	0.270	all	0.015	4115

* groups differing at $|t| > 2.13$ (or 1.65 for C, SC)

Relisted blocks had higher nonmatch rates even though there were not many blocks on which to base the comparisons.

Table 15: Nonmatch Rates by Site and Whether Relisted:

SITE	Relisted	rate	*	s. e.	n
C SC	1. No	0.152	2	0.007	17447
	2. Yes	0.278	1	0.038	363
O SC	1. No	0.271	2	0.008	12385
	2. Yes	0.526	1	0.081	366

* groups differing at $|t| > 1.65$

V. Conclusions and Recommendations

The Integrated Coverage Measurement of the Census 2000 Dress Rehearsal yielded many significant differences between rates of persons not matched. This crude statistic of the number of persons missed among census enumerations overestimates the miscount, since it does not take into account erroneous enumerations and census imputations. But nonmatch rates show, without distraction from those other errors, when census enumerations are less likely to be missed. Owners and whites and persons over 30 years old, especially those over 50, have lower nonmatch rates. Among variables used for poststratification, only sex seemed to get weak confirmation of its importance. Several other groups were also found to have lower nonmatch rates: persons not imputed; those reported by household members rather than proxy; nonmovers; those in households small enough to need only one census form; those living in single family dwellings; and those not living in areas where address listing was so confusing that field interviewers had to redo the work. Those findings were confirmation of prior findings or expectations.

It was also useful to see, despite prior findings or expectations, what variable groups were not discriminated by nonmatch rates: type of enumeration area, address style (i.e. the presence or absence of a full street address or of a P.O. box address), whether or not a block was large and thus subsampled, and how much the area's percent of housing units increased since 1990.

Of course most of these variables are merely related to, rather than causes of, nonmatches, and, even if they could be controlled, there is no guarantee that would improve match rates. But in designing and conducting a census, important characteristics should be kept in mind, aiming to continue questionnaire design, interviewer training, and data processing with care and insight aimed at improving estimates. Evaluation of the poststrata variables is part of getting the most precise and accurate dual system estimate. Knowing that proxy respondents generally give more nonmatch data motivates the push for finding a household member whenever possible.

Monitoring trends in how such variables relate to nonmatch rates over the decades is also part of understanding and interpreting census data. A final recommendation then is to follow this research with Census 2000 investigations of similar and additional variables.

VI. References

- Beaghen, M. (1999). "Modeling Initial Phase and Integrated Coverage Measurement Phase Misses in the Census 2000 Dress Rehearsal," Proceedings of the Section on Survey Research Methods, American Statistical Association, to appear.
- Childers, D. (1999). "The Design of the Census 2000 Dress Rehearsal Integrated Coverage Measurement (ICM) " internal memorandum, U.S. Bureau of the Census.
- Fay, R. (1990). "VPLX: Variance Estimates for Complex Samples," Proceedings of the Section on Survey Research Methods, American Statistical Association, 266-271.
- Feldpausch, R. and Childers, D. (1999). "Erroneously Enumerated in the Census 2000 Dress Rehearsal," Proceedings of the Section on Survey Research Methods, American Statistical Association, to appear.
- Games, P. (1971). "Multiple Comparisons of Means," American Educational Research Journal, 8, 531-565.
- Jones, J. and Childers, D. (1999). "Person Duplication in the Census 2000 Dress Rehearsal," Proceedings of the Section on Survey Research Methods, American Statistical Association, to appear.
- Moriarty, C. and Childers, D. (1993). "Analysis of Census Omissions: Preliminary Results," Proceedings of the Section on Survey Research Methods, American Statistical Association, 629-634.
- Schindler, E. (1999). "Iterative Proportional Fitting in the Census 2000 Dress Rehearsal," Proceedings of the Section on Survey Research Methods, American Statistical Association, to appear.
- Waite, P. and Hogan, H. (1998). "Statistical Methodologies for Census 2000," Proceedings of the Section on Government Statistics and Section on Social Statistics, American Statistical Association, 40-49.

SAMPLE DESIGN FOR THE CENSUS 2000 ACCURACY AND COVERAGE EVALUATION

Randal ZuWallack, Matthew Salganik, and Vincent Thomas Mule, Jr., U.S. Census Bureau
Randal ZuWallack, U.S. Census Bureau, Rm 2501, Bldg 2, Washington DC 20233

Keywords: Accuracy and Coverage Evaluation, Census 2000, Stratified Systematic Sampling

Introduction

Every ten years the Census Bureau attempts to enumerate every person living in the United States. Although a complete count is desired, past experience indicates it is virtually unattainable. According to past census evaluations using demographic analysis, the undercount has ranged from 2.8 million in 1980 to 7.5 million in 1940 (Bureau of the Census, 1997). Beginning with the 1950 census, the Census Bureau began conducting post-enumeration evaluations to estimate census coverage. These evaluations took a case by case matching approach to identify people who were missed and those who were counted. More recent evaluations of this type include the 1980 Post-Enumeration Program (PEP) and the 1990 Post-Enumeration Survey (PES). For the PEP, information based primarily on the Current Population Survey was used to estimate people not counted in the census enumeration (Fay, 1988). A second part of the PEP involved selecting a sample of census records to estimate the number of erroneous census enumerations. Improvements were introduced for the 1990 PES. Rather than using information that was not specifically designed for measuring census omissions, a survey was designed with this sole purpose in mind. As was done in 1980, a sample was also selected for estimating erroneous census enumerations.

In the tradition of improving census evaluations, the Census Bureau is conducting the Accuracy and Coverage Evaluation (A.C.E.) following the Census 2000 enumeration. Similar to the PES, the A.C.E. checks the quality of the census in two ways. One is by comparing data from the census to data collected from an independent sample of housing units to estimate the number of people missed. The other is by selecting a

sample of census records to estimate the number of erroneous census enumerations. This information is combined to determine dual system estimates of the total population and many demographic groups, which is then compared to the census results to estimate coverage rates. This paper discusses all phases of the A.C.E. sample design, how the design was effected by the recent Supreme Court decision on sampling for the Census (Department of Commerce v. United States House of Representatives, 1997), and changes made to the design based on an evaluation of the Census 2000 Dress Rehearsal design.

P Sample and E Sample

Because there are two types of coverage errors, missed people and erroneous inclusions, two samples are selected to evaluate census coverage —the population sample (P Sample) and the enumeration sample (E Sample). The P Sample consists of the people living in the housing units designated for A.C.E. interviews. These units are randomly selected from an address list which is compiled independently of the census list for a sample of geographic areas. The list is referred to as the Independent List. The P-sample people are matched back to the census to determine if they were counted or missed. The E Sample consists of people living in a sample of housing units enumerated in the census. The E-sample people are checked to determine whether they were correctly counted in the census, or whether they were erroneously included. Erroneous enumerations include duplicates, fictitious names, people who were born after census day or people who died prior to census day.

Table 1. P Sample and E Sample Comparison

	P Sample	E Sample
Estimates	Omissions	Erroneous Inclusions
Universe	All housing units in US ¹	Census housing units
PSUs	Block Clusters	Block Clusters

The authors are mathematical statisticians in the Decennial Statistical Studies Division of the U.S. Census Bureau. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau Publications. This report is released to inform interested parties of research and to encourage discussion.

¹All housing units in the United States are eligible to be selected except housing units in Remote Alaska.

Block Cluster

The primary sampling units are block clusters, which are one or more geographically contiguous census blocks grouped together. Census blocks are formed by streets, roads, railroads, streams, etc. Forming block clusters involves a complicated hierarchical algorithm involving many rules and constraints. In general, the goal of block clustering is to produce sampling units that average about 30 housing units.

Integrated Coverage Measurement Survey

Until January 25, 1999, when the Supreme Court ruled that statistical sampling could not be used for the House of Representatives reapportionment, the Census Bureau had planned to conduct an Integrated Coverage Measurement (ICM) Survey. The primary goal of the ICM was to produce accurate and reliable direct state estimates, which would then be used for the reapportionment. Preliminary calculations indicated that the ICM allocation may result in coefficients of variation for the Dual System Estimate of approximately 0.5% in all states and standard errors of about 60,000 in the larger states (Schindler, 1998).

The Supreme Court ruling produced a change in the requirements. Direct state estimates were no longer needed for the reapportionment process, and consequently neither was a 750,000 housing unit sample. In contrast to the ICM, which incorporates the information into the population estimates, the A.C.E. results in a second set of estimates which will be used to evaluate the census and potentially for other purposes.

Because the Supreme Court ruling came too late to entirely redesign the sample, we will select an initial sample of block clusters using the ICM design. The independent list will be comprised of the housing units in these selected clusters, called the A.C.E. listing sample. The sample will be reduced during a later process called the A.C.E. Block Cluster Reduction. This has some limitations. The ICM was designed for efficient direct estimates for state total population. The primary goal for A.C.E., however, is to generate reliable demographic group estimates for the purpose of measuring differential coverage. The ICM sample is being selected using proportional allocation within a state. While this might be efficient for total population estimates, it is not efficient for estimating the population of smaller demographic groups. Overall, due to an increased sample size, we expect the reliability to be better for most of the poststrata estimates than the 1990 PES. Also, we expect the state total population estimates to be more reliable than for the 1990 PES.

Stratification and Sort Variables

Historically, coverage rates in the census have varied for many different groups in the population. In 1990, coverage rates were calculated for 357 poststrata identified by region, geographic area, race, Hispanic origin, age, sex, and tenure (own/rent). Although the estimated undercount for the total population was 1.6%, the estimated undercounts for the 357 groups ranged from -8.29% to 21.27% (Thompson, 1992). The poststrata definitions for Census 2000 are currently being researched and thus are not known. However, we are assuming they will be based on similar variables as in 1990 to account for the differential undercount. In order to estimate the coverage rates for several different poststrata with acceptable precision, there must be an adequate amount of sample selected for each of these poststrata. Since the characteristics of people within a block cluster vary, exact sample sizes for these groups are unattainable. However, the variation in the sample sizes for these groups can be improved by grouping similar block clusters together and selecting a systematic sample across these groups. In an attempt to better control the sample sizes from these different groups, block clusters will be classified into categories based on their estimated size, demographic composition, and level of urbanization.

Block clusters will initially be stratified into four mutually exclusive groups within each state: small block clusters (0-2 housing units), medium block clusters (3-79 housing units), large block clusters (80 or more housing units), and American Indian Reservation (AIR) block clusters. These groups will be sampled at different rates during the selection of the A.C.E. listing sample.

Although there will be no differential sampling within these four sampling strata, the clusters will be sorted by several variables in an attempt to sample a diverse set of block clusters. The first sort variable is the American Indian indicator, which has three categories:

- AIR or trustland
- tribal jurisdiction statistical area, Alaska Native Village statistical area or tribal designated statistical area
- all other areas

The second sort variable is the demographic group. Block clusters will be grouped with other block clusters containing similar demographic proportions based on 1990 census data. Assigning this variable to block clusters is described in more detail in the following paragraph. A third variable used for sorting the clusters is the level of urbanization. Each block cluster will be categorized as an urbanized area with 250,000 or more

people, an urbanized area with less than 250,000 people, or a non-urban area. Finally, the clusters will be sorted geographically using county and cluster number.

To aid in selecting a sample that is well represented by the 6 major race/origin groups as well as owners and renters, block clusters will be classified into 12 demographic groups. Although many block clusters tend to have a large proportion of one demographic group, rarely are they entirely composed of only one, thus many clusters may fit well in two or more categories. To ensure that each cluster is assigned to only one group, a hierarchical assignment rule was developed so that when a cluster exceeds the group threshold, it is assigned to that group. These group thresholds were developed by grouping similar 1990 blocks together using a multivariate clustering method². Table 2 lists these threshold values. The order of the hierarchy gives the smaller demographic groups priority over the larger ones and renters priority over owners.

A.C.E. Listing Sample Selection

For each state, a systematic sample is selected for each of the four strata listed in the previous section. In the following paragraphs, the sampling for the medium and large clusters is discussed, followed by the small block clusters and finally the AIR clusters.

As stated earlier, the Census Bureau was preparing to conduct an ICM during the early stages of the sample design. Thus the 25,000 block clusters were allocated to the states to approximately meet the ICM sample requirements, while maintaining a minimum of 300 block clusters per state. Selecting a sample of block clusters within each state results in approximately 2 million housing units to list. The sampling is done in two steps to guard against a listing workload that would be too formidable to complete in time. If the first systematic sample of block clusters results in a workload that is 10% more than the number of housing units allowed for listing, a second systematic sample is drawn from the first to approximately meet the listing constraint. Large block clusters are selected at a higher rate than medium clusters during the A.C.E. listing sample selection. These higher rates coupled with large block subsampling will result in more clusters represented in sample while keeping the total number of designated interviews within budget.

Table 2. Assignment Rule for Census 2000 A.C.E.

Order	Proportion	Threshold
1	Hawaiian and Pacific Islander Renters	0.10
2	Hawaiian and Pacific Islander Owners	0.10
3	American Indian and Alaska Native Renters	0.10
4	American Indian and Alaska Native Owners	0.10
5	Asian Renters	0.20
6	Asian Owners	0.20
7	Hispanic Renters	0.20
8	Hispanic Owners	0.20
9	Black Renters	0.25
10	Black Owners	0.25
11	White and other Renters	0.30
12	White and other Owners	all others

Small block clusters are generally sampled at a lower rate than both medium and large clusters. This is due to cost considerations which are further explained in a later section. These lower sampling rates cause some small cluster to have high weights, which may disproportionately affect the dual system estimates. In an attempt to avoid the problems associated with the high weights we will initially sample 5,000 small block clusters. Using information about these 5,000 clusters we will attempt to target potential problem clusters in the subsampling operation which will reduce the number of small clusters in sample. These initial 5,000 small clusters were allocated to states proportionately to their projected number of housing units in small blocks. This allocation was bounded by two constraints – a 20 block cluster minimum and a minimum expected sampling rate of 1 in 1000.

To ensure sufficient sample for calculating accurate undercount rates for American Indians on reservations, 355 block clusters will be selected from the block clusters on AIR nationwide. Small block clusters on AIR will not be included in this 355 block clusters. These clusters will be eligible for selection in the small cluster stratum. These 355 clusters were allocated to 26 states proportional to the 1990 population of American

²PROC FASTCLUS in SAS uses a multivariate clustering technique called nearest centroid sorting. For details, refer to pages 824-850 of the SAS/STAT User's Guide, Volume 1, Version 6, Fourth Edition.

Indians on reservations. Ten states contained AIR clusters with little or no American Indian population. These clusters are not included in an AIR stratum, but instead are eligible for selection in the other strata. The remaining 14 states and the District of Columbia contain no block clusters on AIR.

A.C.E. Block Cluster Reduction

As previously stated, the ICM sample will be reduced via the A.C.E. Block Cluster Reduction. This process is the first of three operations that will reduce the 2 million housing units listed down to approximately 300,000 housing units, which is nearly twice the sample size of the 1990 Post-Enumeration Survey (PES). The other two operations are described in the sections that follow. The sample was allocated to the states and the District of Columbia proportional to state population, with a minimum of 1,800 housing units designated for interview per state. The reduction will possibly have variable sampling rates within each state based on race, ethnicity and tenure classification of the block clusters. This differential sampling will help to provide sufficient sample sizes for providing estimates for several different poststrata. In order to provide sample for reliable AIR estimates, the AIR block clusters will not be reduced.

Small Block Cluster Subsampling

Small block clusters, those with between 0 and 2 housing units, get special attention in the A.C.E. These clusters have only a few housing units and are not a cost-effective workload for interviewing and follow-up operations. In order to wisely use our fixed resources we will sample small clusters at a lower rate than both medium and large clusters. Because of these uneven sampling rates the people in small clusters will have high weights. These high weights can disproportionately affect the dual system estimates. In 1990 only about 2.4% of the P sample people and 1.7% of the E sample people lived in small clusters. Yet these clusters contributed almost 10% to the net undercount and 15% to the estimated variance (Fay, 1998). In an attempt to improve our estimates we have developed a special design component to deal with small clusters.

Initially we will select 5,000 small clusters that will be a part of the A.C.E. address listing operation. Then through the small cluster subsampling operation we will reduce the number of small clusters in sample while at the same time attempting to achieve two other goals. First, we would like to prevent any small clusters from having weights that are extremely high compared to other clusters in the sample. Second we would like to limit the weights on the few clusters which we expected to be

small, but turned out to be larger. Both of these goals would help to reduce the variance of the Dual System Estimator.

To achieve these goals we will use differential subsampling where the subsampling rates are based on the number of housing units on the Independent Listing and the number of housing units on the Census List. We are in the process of determining the methodology for attaining both goals.

Large Block Cluster Subsampling

Large block cluster subsampling is the final stage in selecting the housing units that are designated for an A.C.E. interview. The underlying concept of large block subsampling is to select a wide range of clusters, while still remaining within the budgeted number of housing units for interview. Assuming that people within a cluster are similar, interviewing all of them is not the most efficient use of resources. Instead, interviewing a smaller piece of several different clusters should provide a more geographically diverse sample.

This stage involves selecting a portion of each block cluster containing 80 or more housing units³. Housing units are selected by dividing each large cluster into segments of adjacent housing units, that differ by no more than one housing unit. Then, a sample of segments is selected by taking one systematic sample across all large clusters in a state. All housing units in the selected segments are designated for A.C.E. interview. The sampling rate is determined so that the number of units selected for interview in large clusters added to the number selected in non-large clusters is approximately equal to the interviewing budget. In other words, since all housing units in non-large clusters are designated for interview, the difference between this number and the budgeted number of interviews is the target number of designated interviews from the large clusters.

E Sample Identification

Once the housing units have been selected for A.C.E. interview the next operation is to select the housing units that are in the E Sample. The information gathered from these housing units will be used to estimate the number of erroneous inclusions in the census. Although an overlapping P Sample and E Sample is not necessary, it is more cost efficient. If the E Sample includes many of the same people we can use the

³Clusters on American Indian Reservation are not subject to Large Block Cluster Subsampling.

information from the P-sample interview to determine whether they were correctly enumerated and thus do not require a follow-up visit.

In an attempt to create overlapping samples, and thus save money, we will map the block clusters and segments of block clusters that are used to select the P Sample onto the census address list. If this step yields any cluster which will require more than 80 follow-up interviews, the E-sample housing units in these clusters will be subsampled.

Changes from Census 2000 Dress Rehearsal

In 1998, the Census Bureau conducted a Dress Rehearsal to refine the Census 2000 operations. The Dress Rehearsal revealed a few areas in the sample design that needed improvement. Many of the changes were minor operational details, but there are a few enhancements worth noting, two of which involve the treatment of small blocks.

The first change involves the formation of block clusters. Small blocks were not clustered with their neighbors for the Dress Rehearsal. Under certain conditions in 2000, small blocks are clustered with their neighbors. This reduces the total number of small clusters and thus reduces their weights. Overall, this change reduced the number of small clusters by about 65%, from 2,968,956 to 1,029,185. Under the new clustering procedure the initial weights for housing units in small clusters vary from 25 to 632 with an average of 221. Had improvements not been made, they would have ranged from 56 to 1,010 with an average of 588. Figure 1 shows the weight distributions of the 50 states, the District of Columbia, and Puerto Rico using both methods.

Also different in the Dress Rehearsal is the allocation of small clusters to states. In the Dress Rehearsal small clusters were allocated proportionately to the number of medium and large sample clusters in each site. This methodology is inefficient since many states have a large population but very little of it is contributed by small blocks whereas other states have a higher percentage of their population in small blocks. To account for this, in 2000 the small clusters were allocated proportional to the number of housing units projected in small clusters. This generally benefits states with larger proportions of the population residing in small clusters. The two allocations are listed in Table 3 for the states with the five highest and five lowest proportions of the population residing in small blocks.

Much of the A.C.E. operational planning was based on 1990 census data. For instance, the estimated number of housing units for creating the Independent List for each state was estimated based on 1990 information.

Since these numbers were then used for renting office space and hiring staff in different areas of the country, exceeding these numbers may pose workload problems. Thus, these estimates became the listing constraints. To help keep the listing close to the listing constraints, two adjustments were built into the design. The first involves an adjustment prior to selecting a sample which is based on expected values. If it appears the listing would be too much based on the preliminary sampling rate, then the sampling rate was decreased. The second adjustment comes in the form of a two step sample. If the clusters selected during the first step surpass the listing constraint, a second sample from the first sample is selected. Without these two procedures, the listing would have surpassed the constraints by over 7.5 percent.

As can be seen by the sampling of changes listed in the above paragraphs, the A.C.E. sample design is continuously being updated and improved. Although there are still details to develop, such as the sampling rates for the small block subsampling and the possible strata for A.C.E. reduction, the framework is in place to provide reliable estimates of census coverage.

Table 3. Initial Small Block Cluster Weights for Selected States

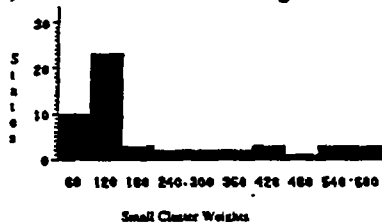
State	Percent 1990 Hus in Small Blocks	Dress Rehearsal Method Weight	Census 2000 Method Weight
North Dakota	11.67%	299	148
South Dakota	9.14%	246	139
Nebraska	5.47%	222	94
Kansas	4.64%	365	113
Wyoming	3.46%	529	617
Rhode Island	0.37%	11	41
New Jersey	0.32%	92	218
California	0.29%	156	467
Hawaii	0.24%	102	306
DC	0.06%	6	25

Figure 1. Frequency of Small Cluster Weights

(a) Dress Rehearsal Clustering



(b) Census 2000 Clustering



References

Bureau of the Census. (1997). Report to Congress: Plan for Census 2000. Washington, D.C.: Bureau of the Census.

Department of Commerce v. United States House of Representatives, No. 98-404 (U.S. filed Jan. 25, 1999).

Fay, R.E. (1988), "Evaluation of Census Coverage from the 1980 Post Enumeration Program (PEP): Census Omissions as Measured by the P Sample", Census Bureau Memorandum, March 10, 1988.

Fay, R. E. (1998), "Small Blocks in the 1990 PES", Census Bureau Memorandum, August 1998 (DRAFT).

SAS Institute Inc., *SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 1*, Cary, NC: SAS Institute Inc., 1989. 943 pp.

Schindler, E. (1998), "Allocation of the ICM Sample to the States for Census 2000," *Proceedings of Survey Research Methods Section, American Statistical Association*, Alexandria, VA, American Statistical Association, to appear.

Thompson, J. (1992), "CAPE Processing Results", Census Bureau Memorandum, March 20, 1992.